GOODNESS-OF-FIT TESTING USING CROSS-VALIDATION BAYES FACTORS

A Dissertation

by

MATTHEW R. MALLOURE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Jeffrey D. Hart
Committee Members,	Thomas Wehrly
	Suojin Wang
	Ximing Wu
Head of Department,	Valen Johnson

December 2017

Major Subject: Statistics

ABSTRACT

Statistical methods for selecting between two competing models have a long and storied history from both the frequentist and Bayesian perspectives. That being said, there are known limitations that exist when using frequentist tests based on P-values for model selection. Therefore, we prefer to take a Bayesian approach to model selection that utilizes Bayes factors. In this research, we consider two different model selection problems: multivariate nonparametric goodness-of-fit and comparing two parametric models. For both problems, we propose intuitive and computationally simple model selection methods that take advantage of data splitting and cross-validation Bayes factors.

Bayesian multivariate nonparametric goodness-of-fit is a difficult problem. The alternative model often requires an infinite-dimensional prior distribution that makes computation of the marginal likelihood complex. By applying data splitting, we are able to form a nonparametric alternative model using the familiar multivariate kernel density estimate and compute a cross-validation Bayes factor very easily.

As for comparing two parametric models (either nested or non-nested), difficulties can arise when formulating prior distributions or approximating marginal likelihoods for either model. We can avoid both of these concerns by computing a *prior-free* cross-validation Bayes factor by using data splitting. These Bayes factors depend solely on computing maximum likelihood estimates and evaluating likelihood functions.

In both scenarios, we show that our cross-validation Bayes factors are consistent at an *exponential rate, regardless of which hypothesis is true*. This includes the traditionally difficult case where the smaller of two nested parametric models is true. We also provide numerous simulation studies and real data analyses to explore performance and practical application of these methods.

DEDICATION

I dedicate this to Yeni, my parents (John and Suzy),

and my sisters (Lisa, Karen, and Kristen).

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to the following individuals for making this dissertation and my doctoral studies possible.

First, I would like to thank Dr. Jeffrey D. Hart for serving as my advisor and allowing me the opportunity to research cross-validation Bayes factors. Dr. Hart's enthusiasm for research and dedication to teaching his students truly made this an enjoyable experience. This dissertation would not have been possible without his guidance, wisdom, meticulous attention to detail, and seemingly infinite patience.

Next, I would like to thank Dr. Thomas Wehrly, Dr. Suojin Wang, and Dr. Ximing Wu for serving as members of my committee and offering their questions, comments, and suggestions along the way.

I am especially grateful to Dr. Michael Longnecker for always having the time to answer questions, give teaching advice, provide needed motivation, see how things were going, and of course talk about everything Michigan over my entire career at Texas A&M.

I would also like to thank all of my professors and classmates whether it be at Grand Valley State or here at Texas A&M, for helping me discover my passion for Statistics and for fostering my education over the last 11 years.

Last, but certainly not least, I am extremely grateful for the love, continued support, and unwavering belief in me expressed by Yeni and my family. I can finally answer your question, my research project is now complete!

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Jeffrey D. Hart [advisor], Professor Thomas Wehrly and Professor Suojin Wang of the Department of Statistics and Professor Ximing Wu of the Department of Agricultural Economics.

The gene expression microarray data analyzed in Chapter 4 were collected by Professor Robert S. Chapkin and associates of the Department of Nutrition and Food Science. A larger analysis of these data can be found in the article by Davidson et al. (2004). The other three data sets can all be found in publicly available repositories. Specifically, the kevlar data (Chapter 2) are found in Andrews and Herzberg (1985), the Academic Performance Index data (Chapter 4) are from the *survey* package in R (Lumley, 2017), and civil engineering data (Chapter 5) are posted on the University of California at Irvine Machine Learning Repository (Lichman, 2013).

The research in Chapter 5 on the parametric CVBF method has been submitted for publication (co-authored with Professor Jeffrey D. Hart).

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a Graduate Merit Fellowship from Texas A&M University and a state funded (Technology) Teaching Assistantship from the Department of Statistics.

NOMENCLATURE

API	Academic Performance Index
BF	Bayes Factor
CVBF	Cross-Validation Bayes Factor
CVBF_K	Kernel Cross-Validation Bayes Factor
$\text{CVBF}_K(\mathcal{A})$	Kernel Cross-Validation Bayes Factor for bandwidth matrix class $\mathcal{A}$
CVBF_P	Parametric Cross-Validation Bayes Factor
CVWE	Cross-Validation Weight of Evidence
DP	Dirichlet Process
DPM	Dirichlet Process Mixture
EDF	Empirical Distribution Function
HJ	Hjort-Jones Estimator
IBF	Intrinsic Bayes Factor
iid	Independent and Identically Distributed
IQR	Interquartile Range
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimate
UIR	Unit Information Reference

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION AND LITERATURE REVIEW

Statistical methods are often derived based on the assumption that the data follow a specific parametric distribution. For instance, methods such as t-tests, linear regression, ANOVA, MANOVA, principal components analysis, and linear discriminant analysis require the data to follow either a univariate or multivariate normal distribution (Korkmaz et al., 2016). While the normal distribution is by far the most prevalent model used, there are situations where distributions such as the chi-square, log-normal, exponential, or Poisson distributions need to be assumed. For example, Rayner and Best (1989) mention that "Safety limits for extreme rainfall used by hydrologists involved in flood control may assume a lognormal distribution", as well as, "Estimates of bacteria in sewage may be based on an exponential distribution". Regardless of which parametric distribution is assumed, when applying a statistical method in practice, the validity of the conclusions will depend on how well the necessary probability model fits the observed data. Therefore, in order to perform a valid statistical analysis, it is of paramount importance to develop statistical methods, known as goodness-of-fit tests, to verify that the underlying data model meets the necessary assumptions. As quoted in Rayner and Best (1989), H.J. David defined a goodness-of-fit test as "... a statistical test of a hypothesis that the sample population is distributed in a specific way" and Oscar Kempthorne coined goodness-of-fit tests as the "classical problem of statistical inference". Prior to the development of goodness-of-fit tests, the only way to assess distributional assumptions was visually, which can only be done feasibly in fewer than three dimensions.

In their most general form, goodness-of-fit tests compare a parametric model to a non-parametric model. The parametric model in this case is a density function (univariate or multivariate) that is indexed by a finite set of unknown parameters. The challenge

facing statisticians in developing these tests lies in forming the nonparametric alternative. Since by definition, a nonparametric model assumes nothing regarding its functional form, the model is infinite-dimensional. Therefore, while defining the parametric model and specifying its parameters is easy, finding a suitable nonparametric model over a high-dimensional function space is often difficult.

1.1 History of Goodness-of-Fit Testing

Goodness-of-fit tests have a long and storied history dating back to 1900 and are still studied extensively to this day. Over the course of the first 100 years of study, goodness-of-fit tests were mainly approached from the frequentist perspective due to computational ease. It was not until the mid 1990s that we saw the first practical application of Bayesian methods to test distributional fit in the statistical literature. In this section, we will explore the history of multivariate goodness-of-fit testing by pointing out the more noteworthy advances from both the frequentist and Bayesian paradigms. As we will see in both paradigms, the majority of the earliest goodness-of-fit tests applied only to univariate data. However, over time, the need for multivariate tests became apparent and the natural approach was to either directly or indirectly extend the univariate tests to multivariate data. In fact, the current thesis extends a univariate approach to multivariate data. Therefore, the history included in this section will not only include the basic details of the main multivariate techniques, but also their respective univariate foundations. With two different approaches to the same problem, it is also important to include the logic behind our decision to take the Bayesian viewpoint. Thus, this section also includes a comparison of frequentist and Bayesian hypothesis testing.

1.1.1 Frequentist Tests

The first goodness-of-fit test found in the statistical literature is Karl Pearson's chi-squared test, published in 1900 (Pearson, 1900). Even though this test is over 100 years

old, it is still taught in every introductory statistics course and used in practical statistical analyses daily. In fact, this seminal work was so monumental in the field of statistics that a conference on *Goodness-of-Fit Tests and Model Validity*, was held in Paris, France in May of 2000 to commemorate its 100 year anniversary (Huber-Carol et al., 2002). As summarized by C.R. Rao (Huber-Carol et al., 2002), the chi-squared test is ideally suited for qualitative data in the form of frequencies for a finite number s of natural categories. In order to compare the fit of the observed data to an assumed discrete probability model (multinomial, Poisson, etc.) we test the hypotheses,

$$H_0 : \pi_i = \pi_i(\theta),\ i = 1, \ldots, s$$
$$H_1 : \pi_i \neq \pi_i(\theta),\ i = 1, \ldots, s$$

where the probability for category i is a function of the completely specified parameter vector θ. To test these hypotheses, Pearson (1900) computed the chi-square statistic,

$$\chi^2 = \sum_{i=1}^{s} \frac{(O_i - E_i)^2}{E_i}, \tag{1.1}$$

which compares the observed frequencies $(O_i = np_i)$ from the data to the expected frequencies $(E_i = n\pi_i(\theta))$ under the assumed model for each of the s categories. The χ^2 statistic is a dissimilarity measure such that smaller values indicate that the observed data are more consistent with the assumed model ($\chi^2 = 0$ indicates a perfect match). Asymptotically, the χ^2 statistic follows a χ^2_{s-1} distribution and tail probabilities can be used to make conclusions about how well the parametric model fits the data.

Pearson's goodness-of-fit test is a specific form of a more general class of goodness-of-fit tests defined by Neyman (1937) known as "smooth" goodness-of-fit tests. "Smooth" refers to departures from the null model that are based on the first four central moments of

the distribution. To carry out Neyman's tests, the null hypothesis is based on applying the probability integral transform to the specified null probability density function $f(x)$ (with cdf $F(x)$) so that $H_0 : Y = F(x) \sim U(0,1)$. The smooth alternative distribution of order k has the form,

$$g_k(y;\theta) = C(\theta)\exp\Big[\sum_{i=1}^{k}\theta_i\pi_i(y)\Big],$$

where $\theta^T = (\theta_1, \cdots, \theta_k)$ is a vector of parameters, $C(\theta)$ is the normalizing constant, and $\pi_i(y)$'s are orthonormal polynomials. The alternative model is considered an extended model since when θ is the zero vector, it reduces to the null model. The resulting Neyman test statistic is

$$\Psi_k^2 = \sum_{i=1}^{k}\Big[\sum_{j=1}^{n}\pi_i(Y_j)/\sqrt{n}\Big]^2, \tag{1.2}$$

which asymptotically follows a χ_k^2 distribution. Neyman's "smooth" goodness-of-fit tests can be extended to a wide variety of scenarios including both simple or composite hypotheses and discrete or continuous models as detailed further in Rayner and Best (1989).

Notice that in Neyman's "smooth" goodness-of-fit tests, the null model is embedded in the alternative model. Thus, the next natural advancement in nonparametric goodness-of-fit testing looked to test the fit of a parametric model $F(\cdot|\theta)$ versus a nonparametric estimate of the true distribution function without this embedding. The most basic nonparametric estimate of a distribution function, $F(x)$, is the empirical distribution function (EDF), $F_n(x) = n^{-1}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$. In order to conduct these tests, researchers considered using dissimilarity measures between F_n and $F(\cdot|\theta)$. Three well known methods based on the EDFs that are still used today are the, Cramér-von Mises, Kolmogorov-Smirnov, and Anderson-Darling tests (D'Agostino and Stephens, 1986). The respective distance measures used in the three tests are provided in Table 1.1. Besides being based on the EDF, these three methods also have the commonalities that they can be used to

Goodness-of-Fit Test	Distance Measure
Cramér-von Mises	$Q = n\int_{-\infty}^{\infty}[F_n(x) - F(x\|\theta)]^2 \mathrm{d}F(x\|\theta)$
Kolmogorov-Smirnov	$D = \sup_x \lvert F_n(x) - F(x\|\theta)\rvert$
Anderson-Darling	$A^2 = n\int_{-\infty}^{\infty} \frac{[F_n(x)-F(x\|\theta)]^2}{F(x\|\theta)(1-F(x\|\theta))}\mathrm{d}F(x\|\theta)$

Table 1.1: Distance measures for univariate goodness-of-fit tests based on empirical distribution functions (D'Agostino and Stephens, 1986).

test goodness-of-fit for any continuous distribution function and their critical values are not based on well-known distribution functions. That being said, there is no one test that has superior performance compared to the others in all situations (presence of outliers, influential points, skewness, heavy/light tails, etc.). In fact, when it comes to specifically testing for univariate normality, the Shapiro-Wilk test is superior (and preferred) to those in Table 1.1. The test statistic for the Shapiro-Wilk test is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{1.3}$$

where x_i is the i-th order statistic and the constants $\mathbf{a} = (\mathbf{a_1}, \ldots, \mathbf{a_n}) = (\mathbf{c'V^{-1}c})^{-1/2}(\mathbf{c'V^{-1}})$ are a function of the expected values ($\mathbf{c}$) and covariance matrix ($\mathbf{V}$) of independent and identically distributed (iid) standard normal random variables (D'Agostino and Stephens, 1986). The four univariate tests mentioned here represent the current gold standard for distributional goodness-of-fit in statistics today, but of course are only a snapshot of the many goodness-of-fit tests in the literature.

So far, the frequentist nonparametric goodness-of-fit tests provided are only applicable to univariate data. As is common with many multivariate methods, once the univariate ver-

sion is well understood, the next natural step is to extend the one-dimensional methods to d-dimensional scenarios. This is precisely what took place in the goodness-of-fit literature for the tests based on the EDF. Some authors indirectly tested multivariate goodness-of-fit by first transforming the multivariate data to scalar data and then applying the univariate methods described above. For instance, Malkovich and Afifi (1973) note that if $X_1, X_2, \ldots, X_n \sim N_d(\mu, \mathbf{\Sigma})$ then asymptotically $Y_i = (X_i - \bar{X}_n)^T \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1} (X_i - \bar{X}_n) \sim \chi_p^2$ and test goodness-of-fit by applying both the Kolmogorov-Smirnov and Cram'er-von Mises tests to the Y_i's. Also, Hawkins (1981) tests multivariate normality by computing a P-value from an F-distribution based on the squared Mahalanobis distance (from the mean) for each observation and then applies an Anderson-Darling test to test for uniformity of the P-values. Finally, Royston (1982) and Villasenor Alva and González Estrada (2009) apply the Shapiro-Wilk test to test multivariate normality by computing the Shapiro-Wilk test statistic for each univariate marginal distribution from the centered and rescaled data. Royston then centers and rescales the d test statistics, computes a weighted average using the normal cumulative distribution function, and uses its respective asymptotic chi-squared distribution to find a P-value (Mecklin and Mundfrom, 2004). Villasenor Alva and González Estrada (2009) simply take the arithmetic average of the Shapiro-Wilk test statistics and numerically compute P-values via Monte Carlo simulation. Each of these authors found creative ways to apply univariate goodness-of-fit tests to multivariate data to assess multivariate goodness-of-fit.

Other authors derived analogous goodness-of-fit tests to the univariate ones that apply directly to multivariate data using the multivariate EDF $F_n(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(X_i \leq \mathbf{x})$, where $X_1, \ldots, X_n \in \mathbb{R}^d$ constitute a random sample from F. Justel et al. (1997) derived a multivariate version of the Kolmogorov-Smirnov statistic for testing $H_0 : F = F_0$, where F_0 is completely specified. The natural test statistic in the multivariate setting would

simply be

$$K = \sup_{\mathbf{x} \in \mathbb{R}^d} |F_n(\mathbf{x}) - F_0(\mathbf{x})|,$$

the largest absolute difference between the null distribution function and the EDF. However, unlike the univariate Kolmogorov-Smirnov statistic, K is not distribution-free. In order to derive a distribution-free statistic, the authors cite the following result of Rosenblatt (1952). Let $Y \in \mathbb{R}^d$ be a random vector with joint density

$$f(y_1, \ldots, y_d) = f_1(y_1) f_2(y_2|y_1) \cdots f_d(y_d|y_1, \ldots, y_{d-1})$$

and define the transformation $U = T(Y)$ by

$$\begin{aligned} U_1 &= F_1(Y_1) \\ U_l &= F_l(Y_l|Y_1, \ldots, Y_{d-1}), \quad l = 2, \ldots, d. \end{aligned}$$

Then, $U_1, \ldots, U_d \overset{iid}{\sim}$ uniform$[0, 1]$. Applying this multivariate probability integral transform to the observed data, the test statistic for testing d-dimensional uniformity is given by

$$D = \sup_{\mathbf{u} \in [0,1]^d} |G_n(\mathbf{u}) - \prod_{l=1}^{d} u_l|,$$

where $G_n(\cdot)$ is the EDF of the transformed data. However, due to the sequential nature of the transformation, D is not invariant to permutations of the coordinates. Therefore, the multivariate Kolmogorov-Smirnov statistic D_d^{KS} is the maximum of D over all possible permutations of the coordinates. Chiu and Liu (2009) extended these ideas to provide a multivariate version of the Cramér-von Mises statistic with the following form:

$$\int_{[0,1)^d} \left| G_n(\mathbf{u}) - \prod_{l=1}^{d} u_l \right|^2 d\mathbf{u}.$$

Notice in both cases, the similarity between the multivariate and univariate versions of the Kolmogorov-Smirnov and Cramér-von Mises test statistics. However, implementing both of these tests can become computationally demanding for d of any size due to requiring all $d!$ permutations of the coordinates.

The multivariate normal distribution is the most common parametric model in a goodness-of-fit test, so the final two tests that we will talk about, which are not direct extensions of univariate methods, are specifically for testing normality. Mardia's test (Mardia, 1970) is one of the first tests of multivariate normality and is based on multivariate measures of skewness and kurtosis, respectively given by,

$$b_{1,d} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} [(X_i - \bar{X})^T S^{-1} (X_j - \bar{X})]^3$$
$$b_{2,d} = n^{-1} \sum_{i=1}^{n} [(X_i - \bar{X})^T S^{-1} (X_i - \bar{X})]^2.$$

Asymptotically, $\frac{nb_{1,d}}{6} \sim \chi^2_{(d(d+1)(d+2))/6}$ and $b_{2,d} \sim N(d(d+2), 8d(d+2)/n)$. Each of these two statistics and their respective P-values are simple to compute, making them very attractive in practice. The Henze-Zirkler test (Korkmaz et al., 2016) is based on a non-negative functional that measures the distance between the empirical and parametric distribution functions. The formula for the test statistic is given by

$$HZ = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \exp\left(-\frac{\beta^2}{2} D_{ij}\right) - 2(1+\beta^2)^{-d/2} \sum_{i=1}^{n} \exp\left(-\frac{\beta^2 D_i}{2(1+\beta^2)}\right) + n(1+2\beta^2)^{-d/2},$$

where $\beta = 2^{-1/2} 4^{-1/(d+4)} (2(2d+1))^{1/(d+4)}$, D_{ij} and D_i are the squared Mahalanobis distances between the ith and jth observations and the ith observation and the mean, respectively. For normal data, HZ follows a lognormal distribution, which is also easy to work with. Since both Mardia's test and the Henze-Zirkler test are easy to implement in

practice, they are arguably the most common approaches to testing multivariate normality.

Since no one test is preferable the best in all situations, research for univariate and multivariate goodness-of-fit tests is still on-going and there are many more frequentist methods that are not covered here. D'Agostino and Stephens (1986) provide a more thorough survey of goodness-of-fit tests for a variety of parametric models. Regarding tests specifically for normality, Thode (2002) provides dozens of different tests primarily focused on univariate data. For a summary of testing multivariate normality, see Mecklin and Mundfrom (2004). Finally, for a more theoretical look at goodness-of-fit tests (and hypothesis tests for that matter) see Lehmann and Romano (2005).

1.1.2 Comparison of Frequentist and Bayesian Hypothesis Testing

Based on the discussion in the previous subsection, it is clear that the distributional goodness-of-fit problem has been well studied from the frequentist perspective. Before delving into the history of Bayesian approaches to goodness-of-fit testing, the following comparison of frequentist and Bayesian hypothesis testing provides an explanation as to why we prefer using a Bayesian approach. Under the frequentist goodness-of-fit testing framework, the hypotheses are typically defined as

$$H_0 : Y_1, \ldots, Y_n \sim f(\cdot|\theta \in \Theta)$$
$$H_1 : Y_1, \ldots, Y_n \nsim f(\cdot|\theta \in \Theta)$$

where $Y_1, \ldots, Y_n$ comprise the random sample and $f(\cdot|\theta)$ is the parametric model of interest, indexed by parameter vector θ in parameter space Θ. Now, when carrying out the test, we either "reject" or "fail to reject" the null hypothesis. Therefore, for small P-values, we "reject" the null model, but this tells us very little regarding which model we should consider next since the alternative model is often vague (potentially infinite-dimensional).

On the other hand, for large P-values (those larger than the size of interest) we "fail to reject" the null model, thus considering it a plausible model, but we cannot "accept" the null model as truth. Notice also that these conclusions about the null model are based on the usual P-value. Often, the typical cutoff value for rejecting the null is $\alpha = .05$. However, for $P = .05$, Delampady and Berger (1990) show that the conclusion made based on a given P-value can be misleading and more often than not it provides more evidence against the null than exists. This problem of comparing the Bayesian and frequentist tests to illustrate the contradicting amount of evidence for/against the null model has been studied by many authors, namely: Lindley (1957), Berger and Delampady (1987), and Berger and Sellke (1987). In more recent research, Johnson (2013) argues that carrying out hypothesis tests of size $\alpha = .05$ is inappropriate, especially for large data sets. One can easily construct simple examples showing that, as $n \to \infty$, the level of significance should tend to 0. Therefore, taking a frequentist approach to goodness-of-fit testing is less than ideal.

Suppose we now consider goodness-of-fit tests from the Bayesian perspective. In the most simple Bayesian hypothesis testing framework, we require a well-defined probability model for both the null and alternative hypotheses. Therefore, the hypotheses often take the form (taken from Verdinelli and Wasserman (1998)),

$$H_0 : \mathbf{Y} \sim \mathcal{F}_0 = \{f(\cdot|\theta_0); \theta_0 \in \Omega_0\}$$
$$H_1 : \mathbf{Y} \sim \mathcal{F}_1 = \{g(\cdot|\theta_1); \theta_1 \in \Omega_1\}.$$

Given these two models, we compute the Bayes factor in favor of the alternative model

$$\text{BF} = \frac{\int \left[\prod_{i=1}^{n} g(Y_i|\theta_1)\right] p(\theta_1) d\theta_1}{\int \left[\prod_{i=1}^{n} f(Y_i|\theta_0)\right] \pi(\theta_0) d\theta_0},$$

which is the ratio of marginal likelihoods. According to Lavine and Schervish (1999), the

Bayes factor measures the change in the odds in favor of the alternative model from the prior to the posterior (after observing the data). Furthermore, when the prior odds ratio of the two models is 1, the Bayes factor *is* the posterior odds. This can be seen rather easily by applying Bayes Theorem to the posterior probabilities of each model to show that,

$$\frac{P(\mathcal{F}_1|\mathbf{Y})}{P(\mathcal{F}_0|\mathbf{Y})} = \frac{P(\mathbf{Y}|\mathcal{F}_1)}{P(\mathbf{Y}|\mathcal{F}_0)} \frac{P(\mathcal{F}_1)}{P(\mathcal{F}_0)}.$$

For a more detailed description of Bayes factors see Kass and Raftery (1995) and for Bayesian hypothesis testing see Gelman et al. (2014). The beauty of using Bayes factors for model comparison is that regardless of the hypothesized models, the set of critical values indicating the amount of evidence in favor of the alternative (or null) model is the same. The two most common sets of critical values can be found in Appendix B of Jeffreys (1961) and in Kass and Raftery (1995). For the purposes of this research, we will consider the scale in Kass and Raftery (1995) given in Table 1.2. This is unlike frequentist

BF	$\log$ BF	Evidence for $\mathcal{F}_1$
1 to 3	0 to 1.1	Not worth more than a bare mention
3 to 20	1.1 to 3	Positive
20 to 150	3 to 5	Strong
> 150	> 5	Very strong

Table 1.2: Amount of evidence in favor of the alternative model for varying values of a Bayes factor (Kass and Raftery, 1995)

goodness-of-fit tests where the critical values depend on the asymptotic distribution of the

test statistic. Each of the tests provided in Subsection 1.1.1 utilize a different test statistic and thus a different asymptotic distribution. In fact, for some tests the null hypothesis is rejected for large values of the test statistic and for others, the null is rejected for small values. The appropriate conclusions regarding the null model are either based on computing tail probabilities analytically or comparing the test statistic to specific critical values. These corresponding P-values do not even quantify how much evidence for/against the null model was obtained from the observed data. In Bayesian hypothesis testing, since the Bayes factor is a ratio, conclusions in favor of the null model can be made using the same scale in Table 1.2 and the reciprocal of $\text{BF} = \text{BF}^{-1}$. This role reversal of H_0 and H_1 is not possible in frequentist testing as the entire testing problem has been fundamentally changed.

One important property of a good Bayesian hypothesis test that utilizes Bayes factors for model comparison is known as Bayes factor consistency and is defined in Definition 1 (Chib and Kuffner, 2016).

Definition 1. *(Bayes Factor Consistency): The Bayes factor defined by BF comparing the alternative model $\mathcal{F}_1$ to the null model $\mathcal{F}_0$ in Bayesian hypothesis testing is consistent if, as $n \to \infty$:*

- *BF* $\to \infty$ $(\log(BF) \to \infty)$ *when $\mathcal{F}_1$ is the true model; and*
- *BF* $\to 0$ $(\log(BF) \to -\infty)$ *when $\mathcal{F}_0$ is the true model.*

As the sample size increases in tests that satisfy Definition 1, the evidence in favor of the true model increases, regardless of which hypothesis is true. This implies that the probability of making a Type I or Type II error tends to 0 as the sample size increases. Certainly, this is what we expect in any hypothesis test, but is not necessarily true in frequentist goodness-of-fit tests as described previously.

A final benefit to Bayesian hypothesis testing is we can now "accept" or "reject" the null model. This means we can determine which one of the two models best fits the data since BF is the updated odds that the observed data were sampled from the alternative model compared to the null model. Of course, even if both models fit the data poorly, one will still be preferred. However, as we will see, our alternative family is defined in such a way that some member of the alternative will be close to the truth in a Kullback-Leibler discrepancy (Kullback and Leibler, 1951) sense.

Certainly, from a philosophical viewpoint, the Bayesian approach to hypothesis testing is far more suitable to testing goodness-of-fit compared to the frequentist approach. The question becomes, why did it take until 1996 for the first practical application of Bayesian hypothesis testing to any goodness-of-fit problem to appear in the literature, when prominent Bayesians were well aware of the deficiencies of frequentist goodness-of-fit tests prior to this time? Remember that the Bayes factor in a Bayesian hypothesis test requires both the null and alternative marginal likelihoods. The null marginal likelihood is a finite dimensional integral over the parameter space, whereas, the alternative marginal likelihood is often a very high (or even infinite) dimensional integral. At least one of these integrals must be computed using either a numerical integration scheme or a Markov Chain Monte Carlo (MCMC) algorithm (see Evans and Swartz (1995) and Robert and Casella (2004) for more details). Therefore, practical implementation of a Bayesian nonparametric goodness-of-fit test was not possible until sufficient computing power and resources were readily available. So prior to this time, frequentist methods were preferred simply because the test statistics and P-values were easy to compute analytically.

1.1.3 Bayesian Tests

The history of Bayesian nonparametric goodness-of-fit tests follows a similar path as the frequentist tests of Subsection 1.1.1 in terms of the sequence of advancements; how-

ever, the research is far less dense. Many of the initial Bayesian goodness-of-fit methods for univariate data are based on Bayesian nonparametric density estimation techniques that utilize Dirichlet processes, Pólya tree processes, and Gaussian processes. These techniques are summarized briefly in Müller and Quintana (2004), in more detail from a theoretical perspective in Ghosh and Ramamoorthi (2003), and in more detail from an applied perspective in Müller et al. (2015).

Carota and Parmigiani (1996) published the first Bayesian nonparametric goodness-of-fit test which utilized a Dirichlet process prior on $\mathcal{F}$, the family of all probability distribution functions F when forming the nonparametric alternative model. The parametric model of interest is denoted by $F_0(\cdot|\underline{\theta})$ with parameter $\underline{\theta} \in \Theta$. In their approach, the data vector $\mathbf{y} = y_{ij}$ is comprised by subsequences $i = 1, \ldots, s$ of lengths $j = 1, \ldots, n_i$. Thus, the alternative model is defined as

$$\begin{aligned}
\mathbf{y}|\underline{F}, \underline{\theta} &\sim \prod_{i=1}^{s} \prod_{j=1}^{n_i} F_i(y_{ij}|\theta_i), \quad \underline{F} = (F_1, \ldots, F_s) \in \mathcal{F} \quad \underline{\theta} = (\theta_1, \ldots, \theta_s) \in \Theta \\
\underline{F}|\underline{\theta} &\sim \prod_{i=1}^{s} \mathcal{D}_i, \quad \mathcal{D}_i \sim \text{Dirichlet process with measure } \alpha_i(\theta_i, y) \\
\underline{\theta} &\sim \mathcal{Q},
\end{aligned}$$

where $\mathcal{Q}$ is the prior distribution function of $\underline{\theta}$. The null model, $\mathcal{F}_0$ is then defined as the class of distributions $\underline{F}_0 = (F_{01}(\cdot|\theta_i), \ldots, F_{0s}(\cdot|\theta_s))$ where $\underline{F}_0$ is the vector of conditional means of the Dirichlet process. Using these two models, the Bayes factor in favor of the null model can be easily computed, but it has one very significant drawback. As Berger and Guglielmi (2001) point out, it is inappropriate to use Dirichlet processes when testing an absolutely continuous null model. In fact, this is verified by Carota and Parmigiani (1996) in Corollary 2, where if no ties exist in the data, (i.e. the data are absolutely continuous), the Bayes factor only depends on the data via the sample size. So, while this first approach

was novel at the time, its use in practice is severely limited.

The next approach to Bayesian goodness-of-fit testing was proposed by Verdinelli and Wasserman (1998) and used a Gaussian process prior in the alternative model. Their approach is very similar in principle to the "smooth" tests of Neyman (1937) in that they embed the parametric null model in an infinite-dimensional exponential family to form the alternative model. The null model is defined by $\mathcal{F} = \{F(\cdot|\theta) : \theta \in \Omega\}$ and a random variable from this model is expressed as $Y = F^{-1}(U|\theta)$ where $U \sim \mathcal{U}(0,1)$ and $\theta \in \Omega$, using the inverse probability integral transform. The alternative model is an extended model defined by the infinite-dimensional exponential family of distributions on $[0,1]$, $\mathcal{G} = \{G(\cdot|\psi) : \psi \in \S\}$. If a random variable Y was from this extended model, then Y could be expressed as $Y = F^{-1}(U|\theta)$ where $U \sim G(\cdot|\psi)$, $\theta \in \Omega$ and $\psi \in S$. When $\psi = \psi_0$, $G(\cdot|\psi_0) = \mathcal{U}(0,1)$ and hence $\mathcal{G} = \mathcal{F}$. The probability densities associated with $\mathcal{G}$ can be written as

$$g(u|\psi) = \exp\left(\sum_{j=1}^{\infty} \psi_j \phi_j(u) - c(\psi)\right),$$

where $\psi = (\psi_1, \psi_2, \ldots)$ are the polynomial coefficients for the rescaled Legendre polynomials $\phi = (\phi_1, \phi_2, \ldots)$ and $c(\psi)$ is the normalizing constant. The unknown parameters, θ and ψ are taken to be independent such that the prior distribution is given by $p(\theta, \psi) = p(\theta)p(\psi)$. Take $p(\theta)$ to be any standard reference prior for θ and $p(\psi) = \prod_{j=1}^{\infty} p(\psi_j|\tau)p(\tau)$ such that each $\psi_j \sim N(0, \tau^2)$ and τ follows a truncated standard normal distribution on $[0, \infty)$. Under this construction, $\sum_{j=1}^{\infty} \psi_j \phi_j(u)$ is a Gaussian process. To use this method in practice, the infinite series must be truncated to a finite number of terms and computation of the Bayes factor requires Metropolis-Hastings algorithms embedded in a Gibbs sampler when calculating the alternative marginal likelihood. The benefit to this approach is its applicability to any absolutely continuous null model for univariate data.

Berger and Guglielmi (2001) offered another approach to goodness-of-fit whereby the

alternative model is based on a mixture of Pólya tree processs. The motivation for using Pólya trees stems from their ability to nonparametrically model continuous densities as well as maintaining objective, noninformative priors for unknown parameters. Also, the Pólya tree process remains flexible as an alternative model since after specifying the mean, there are many free parameters remaining. They argue that this is superior to Dirichlet processes since only one free parameter remains post null specification. The formulation of the alternative model requires embedding the parametric model into the Pólya tree process by choosing the appropriate mixture of Pólya trees (the authors provide two such mixtures that are not included here). Computing the Bayes factor in favor of the null model merely requires a Monte Carlo approximation based on importance sampling, but again there is a significant drawback as noted by Tokdar and Martin (2013). Due to the required binary tree of partitions, this approach does not scale well with increasing dimension and thus is limited to univariate data.

One of the newer Bayesian nonparametric goodness-of-fit tests was published by Tokdar and Martin (2013). Their method is specifically designed to test for normality in any dimension using a non-subjective Dirichlet process mixture of normals as the alternative model. They argue that this approach is superior to those already described because the alternative model only depends on the precision parameter of the Dirichlet process and the Dirichlet process mixture of normals is computationally more efficient to use in any dimension. The setup of their method is rather straightforward in principle. Let $\boldsymbol{X}_{1:n} = (X_1, \ldots, X_n)$, where each $X_i \in \mathbb{R}^d$, denote n independent draws from an unknown d-variate distribution F. We want to test normality of the data, so the null model of interest is $\mathcal{F}_0 = \{F_{\mu,\Lambda} : \mu \in \mathbb{R}^d, \Lambda \in \mathbb{L}_d\}$ where $\Lambda\Lambda'$ is the covariance matrix in Cholesky decomposition form and $\mathbb{L}_d$ is the set of all $d \times d$ lower-triangular matrices. Now, for any (μ, Λ) a Dirichlet process mixture of normals denoted by $\text{DPM}_{\mu,\Lambda}(\alpha, \Psi)$ is the distribution

of the random probability measure

$$\bar{F}_{\mu,\Lambda} = \int \mathrm{N}(\mu + \Lambda u, \Lambda V \Lambda') d\bar{\Psi}(u, V), \quad \text{where } \bar{\Psi} \sim \mathrm{DP}(\alpha, \Psi)$$

Much like the prior distribution for the Berger and Guglielmi (2001) approach, the prior distribution for both models is the same. Here, the right Haar measure is the prior of choice given by $d\pi_H(\mu, \Lambda) = \prod_{j=1}^{d} \Lambda_{jj}^{j-d-1} d\mu d\Lambda$. The difficulty with this approach is calculating the marginal likelihood for the alternative model (the null marginal likelihood is analytically tractable). In order to approximate the integral

$$\int_{\mathbb{R}^d \times \mathbb{L}_d} \int \left[\prod_{i=1}^{n} dF(X_i) \right] d\mathrm{DPM}_{\mu,\Lambda}(F|\alpha, \Psi) d\pi_H(\mu, \Lambda),$$

Tokdar and Martin (2013) recommend computing the inner integral using sequential imputation, and then embed this algorithm in an importance sampling scheme (Basu and Chib, 2003). This is certainly a more complex numerical integration technique compared to previous methods. Also, while this approach can tackle goodness-of-fit in d dimensions, it only applies to testing normality, which is an unfortunate restriction.

There are a few other Bayesian nonparametric goodness-of-fit tests in the literature that will not be described in detail here. First, Conigliani et al. (2000) sought to find a Bayesian alternative to the chi-square goodness-of-fit test for binomial and Poisson data in cases with weak prior information regarding the parameters of the null model and the form of the alternative model. Using fractional Bayes factors (O'Hagan, 1995), since their noninformative priors are improper, they showed via numerous examples that their Bayesian approach was comparable to using the Anderson-Darling statistic. Of course, this approach suffers from the same problem as the chi-square test in that discretization of continuous data is required, which is overly restrictive and causes a loss of information.

Another approach by McVinish et al. (2009) considered an alternative model based on mixtures of triangular distributions. These authors provided a set of sufficient (but not necessary) conditions in which Bayes factor consistency holds and verified that these conditions are met for mixtures of triangular distributions. They argued that showing Bayes factor consistency theoretically for the alternative models listed thus far proves to be very challenging, however consistency does appear to hold in all previous methods. They also claim that mixtures of triangular distributions are easy to work with and estimate smooth density functions well, but details regarding practical implementation of their method are lacking, let alone practical use beyond one dimension.

Certainly, from the Bayesian perspective there are many approaches to testing goodness-of-fit for probability distributions, so why should we consider yet another one? Each of the Bayesian approaches listed have their respective downsides, some of which are: restriction to univariate data, inability to test continuous null models, restriction to testing a normal null model, and complex computation of at least one marginal likelihood in the Bayes factor. Also, the formulation of the alternative model using the partitioning schemes and/or mixtures of various processes is often neither intuitive nor simple. The literature on practical Bayesian goodness-of-fit tests is already sparse to begin with, but none of these tests can be used to test any multivariate parametric null model.

The main method we propose in this research is a simple and intuitive approach for testing multivariate goodness-of-fit for any absolutely continuous parametric null model. It is based on the most recent Bayesian nonparametric goodness-of-fit testing procedure, the kernel cross-validated Bayes factor (CVBF_K) approach proposed by Hart and Choi (2016) (described in much more detail in Chapter 2). In their paper, the authors explore a novel goodness-of-fit approach for univariate data where the alternative model is based on a family of kernel density estimates. Therefore, the alternative model only has one unknown parameter, which means the marginal likelihood can be computed by one-

dimensional numerical integration. Since we know kernel density estimates can be used to estimate multivariate densities, it only seems natural to consider extending the CVBF_K to test multivariate goodness-of-fit.

The kernel CVBF method addresses goodness-of-fit by comparing a parametric model to a nonparametric one using a Bayes factor. We can also use Bayes factors for model comparison where both the null and alternative models are parametric. In a typical Bayesian analysis though, we still need to determine prior distributions for all parameters as well as compute the necessary marginal likelihoods, which can be rather daunting tasks. Therefore, we also propose a secondary method that still uses the idea of cross-validation Bayes factors. However, computing the Bayes factor becomes a trivial task in that we simply evaluate a likelihood ratio. When testing two parametric models, this parametric CVBF (CVBF_P) approach is extremely simple to compute, easy to interpret, and has nice large sample properties under both nested and non-nested models. Of course, this is no longer a nonparametric goodness-of-fit test, but it does make many tests that are difficult using traditional Bayesian methods extremely straightforward.

1.2 Research Layout

The remainder of this research contains the following chapters. Chapter 2 continues the literature review by examining the univariate CVBF_K method of Hart and Choi (2016) in more detail to set the foundation before we consider multivariate data. In Chapter 3, we briefly introduce multivariate kernel density estimates and discuss how the bandwidth matrix will potentially impact the CVBF_K method. In Chapter 4, we formally combine Chapters 2 and 3 to extend the CVBF_K method to multivariate data. Chapter 5 formalizes the parametric CVBF method and Chapter 6 provides a look at future work that combines material from Chapters 1-5.

2. UNIVARIATE CVBF$_K$ METHOD

2.1 General Description

The alternative hypothesis in a typical Bayesian nonparametric goodness-of-fit test is a broad class of nonparametric models (including the null model as a special case) indexed by a large (and sometimes infinite) number of unknown parameters. Looking for a way to simplify this goodness-of-fit testing process, Hart and Choi (2016) consider an alternative model based on a family of univariate kernel density estimators indexed solely on the smoothing parameter, $h > 0$. Kernel density estimates are attractive for use in the alternative model because they are familiar and easy to implement. Also, provided that the true data generating density is smooth, we can assume that at least one of the estimates in the alternative model is close to the true density function. Therefore, regardless of which model the CVBF$_K$ method favors, the resulting model will be a well-defined probability model that fits the observed data well.

One important detail left out in the previous paragraph is that the family of kernel density estimates is only well-defined once data are given. Hence, the Hart and Choi (2016) procedure is called the kernel cross-validated Bayes factor method because we need to use data splitting in order to compute a kernel estimate. For a given random split of the data into a training and validation data set, the Bayes factor is computed on the validation data given the training data. The resulting overall Bayes factor for the test is the geometric average of the individual Bayes factors over numerous random data splits.

2.2 Formal Methodology

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a univariate random sample from some unknown parametric density function g. Suppose we want to test that $g = f(\cdot|\theta)$, where f is a specific density function indexed by parameter vector θ. According to the CVBF$_K$ method, the

hypotheses for this test are written as

$$H_0 : \mathbf{X}^V \sim M_0 = \{f(\cdot|\theta) : \theta \in \Theta\}$$
$$H_1 : \mathbf{X}^V \sim M_1 = \{\hat{f}(\cdot|\mathbf{X}^T, h) : h > 0\}.$$

In the alternative model,

$$\hat{f}(x|\mathbf{X}^T, h) = \frac{1}{mh} \sum_{i=1}^{m} K_1\Big(\frac{x - X_i}{h}\Big), \tag{2.1}$$

is the typical univariate kernel density estimator with kernel function K_1 taken to be a symmetric, unimodal, finite variance density function. The authors recommend using the Gaussian kernel function for its convenient properties and ease of implementation. Also, in order for this alternative model to be well-defined, the data vector $\mathbf{X}$ is randomly split into a training data set, $\mathbf{X}^T = (X_1, X_2, \ldots, X_m)$, and a validation data set, $\mathbf{X}^V = (X_{m+1}, X_{m+2}, \ldots, X_n)$. An appropriate choice of training set size will be discussed later.

With both the null and alternative models well-defined, we can compute a Bayes factor to determine which of the two models best fits the data. The Bayes factor for a single random split in favor of the alternative model is given by

$$\text{BF}_m = \frac{\int_0^\infty \prod_{j=m+1}^{n} \hat{f}(X_j|\mathbf{X}^T, h)p(h)dh}{\int_\Theta \prod_{j=m+1}^{n} f(X_j|\theta)\pi(\theta)d\theta}. \tag{2.2}$$

In order to mitigate the dependence of our conclusions on a given random split, we randomly split the data N times such that $(\boldsymbol{X}_k^T, \boldsymbol{X}_k^V)$ represents the training and validation sets from the kth split $(k = 1, 2, \ldots, N)$, respectively. The resulting $\text{CVBF}_{m,N}$ value is

the geometric mean,

$$\text{CVBF}_{m,N} = \left(\prod_{k=1}^{N} \text{BF}_{m,k} \right)^{1/N},$$

where $\text{BF}_{m,k}$ represents the Bayes factor computed using $(\boldsymbol{X}_k^T, \boldsymbol{X}_k^V)$. In practice, we often consider the weight of evidence, $\log(\text{CVBF}_{m,N})$ when making conclusions for the hypotheses. Therefore, define the cross-validation weight of evidence (CVWE) to be $\text{CVWE}_{m,N} = \log(\text{CVBF}_{m,N})$. For notational simplicity, we may drop the subscripts m and N and simply refer to either a CVBF or CVWE value.

In order to compute the Bayes factor in equation (2.2), we require prior distributions, $\pi(\theta)$ and $p(h)$. Hart and Choi (2016) suggest taking a unit-information, reference (UIR) prior for $\pi(\theta)$ that contains the same amount of information (in terms of Fisher's Information) as one observation from the data and is centered at the observed data. The authors mention that using UIR priors results in Bayes factors that are invariant to location and scale when testing univariate normality. Maybe more importantly, clever choice of the prior distribution (perhaps a (semi-) conjugate prior) under the null model can ease the computational burden on the method provided that closed-form expressions for the marginal likelihood exist.

Deriving an appropriate prior for the smoothing parameter is a bit more complicated. Typically, since the bandwidth in a kernel density estimator acts like a scale parameter, the natural prior to consider first is the scale-invariant, improper prior $p(h) \propto h^{-1}$. However, proper priors must be used in any Bayesian hypothesis test since, when using an improper prior, the Bayes factor is proportional to an arbitrary constant. Therefore, we can find a proper prior by using the intrinsic Bayes factor (IBF) idea proposed by Berger and Perrichi (1996). According to the IBF idea, take the minimal sample size (in this case two observations) so that $L(X_1|X_2, h)p(h) \propto h^{-2}K_1\left(\frac{X_1 - X_2}{h}\right)$ produces a proper posterior distribution. After normalizing, this posterior distribution is used as the prior distribution

for the entire sample. For the Gaussian kernel function the proper posterior distribution has a closed form given by

$$p(h|\beta) = \frac{2\beta}{\sqrt{\pi}h^2} \exp\Big(-\frac{\beta^2}{h^2}\Big), \tag{2.3}$$

where β^2 is a robust estimate of $.5\mathrm{E}\left[(X_1 - X_2)^2\right] = \sigma^2$ calculated from the validation data. Therefore, take $\beta = \mathrm{IQR}(\mathbf{X}^V)/1.35$. In this prior distribution, as $h \to 0$, the prior takes on values nearly 0. In kernel density estimation, as $n \to \infty$, the optimal choice of the bandwidth parameter tends to 0. Therefore, this form of prior is similar in principle to non-local priors of Johnson and Rossell (2010) since it greatly downweights the most plausible values of h under the null model. Thus, in order to conclude in favor of the null model, there must be overwhelming evidence to indicate that the data truly are from the null model. That being said, Hart and Choi (2016) provide a proof that shows the Bayes factor in (2.2) is consistent at an exponential rate under both the null and alternative hypotheses regardless of the form of prior distribution. Typically, the convergence rate is exponential only under the alternative model when testing a parametric null model against a nonparametric alternative model (McVinish et al., 2009), so the kernel CVBF method has improved asymptotic properties compared to other Bayesian nonparametric goodness-of-fit tests.

Unlike the null marginal likelihood that is often analytically tractable, the marginal likelihood under the alternative model must be computed numerically. This is not a significant concern since evaluating the marginal likelihood amounts to a one-dimensional integration problem. Hart and Choi (2016) utilize the *integrate* function in R (R Core Team, 2016); however, there are many efficient numerical integration techniques available, such as Simpson's approximation, Gaussian quadrature, and the Laplace approximation (Davis and Rabinowitz, 2007). We prefer to use the Laplace approximation where applicable

since it is faster, more reliable, and avoids underflow problems since the optimization and evaluation are both on the log scale.

In order to implement the CVBF_K method, we only need to determine the training set size m and the number of random splits of the data N. Theoretically, methods for specifying m and N for a given data set were still open problems in Hart and Choi (2016), however, the authors typically set $N = 100$. For choosing the training set size, one can argue that $.05n < m < \frac{n}{2}$ since the training set needs to have enough observations for the kernel estimate to approximate the observed density function sufficiently well. However, it also should not contain more observations than the number used to evaluate the Bayes factor for a single data split. As seen in Hart and Choi (2016), as m increases, the $\text{CVWE}_{m,N}$ value increases monotonically to 0 under the null model. In contrast, when the alternative model is true, the $\text{CVWE}_{m,N}$ value increases for increasing m until a maximum is reached and then it decreases toward 0. So in order to determine the value of m for a specific sample data set, Hart and Choi (2016) recommend using a scheme called *calibration*, consisting of the following 6 (slightly modified) steps:

1. Carry out the CVBF_K method for numerous training set sizes (every integer between $\lfloor .05n \rfloor$ and $\lceil .5n \rceil$) with a sufficiently large number of random splits, N.

2. Plot a curve of the $\text{CVWE}_{m,N}$ values against the training set sizes. Determine CVWE_{max}, the maximum of the curve, and its corresponding training set size m_{max}.

3. If $\text{CVWE}_{\text{max}} < 0$, then conclude in favor of the null model.

4. If $\text{CVWE}_{\text{max}} > \log(3)$, indicating positive evidence in favor of the alternative model (Kass and Raftery, 1995), carry out the CVBF_K method using m_{max} and N for 500 simulated, independent data sets from the null model. Plot a histogram for these 500 CVWE_0 values.

5. Conclude in favor of the alternative model if $\text{CVWE}_{\text{max}} > \log(3)$ and (nearly) all $\text{CVWE}_0 < 0$ in the histogram.

6. If $0 \leq \text{CVWE}_{\text{max}} < \log(3)$, cautiously favor the null model as there is not enough evidence to favor the alternative model.

This calibration technique has a frequentist flavor to it because we repeatedly sample data from the null model, compute a CVWE value for each data set, and check to see how often these values exceed 0. In the fifth step, it is not guaranteed that all 500 CVWE_0 values will be negative. As long as the CVBF_K method behaves well under the null model, we can reasonably conclude in favor of the alternative model provided that $\text{CVWE}_{\text{max}} > \log(3)$ indicating positive evidence against the null model (Kass and Raftery, 1995).

2.3 Real Data Example: Kevlar Strand Data

In order to see how to implement the CVBF_K method and compare its respective performance to some of its Bayesian nonparametric counterparts, consider the time to failure (or static fatigue) data for each of 100 Kevlar 49 epoxy strands under 80% stress found in the textbook by Andrews and Herzberg (1985). It is hypothesized that the lifetimes $(X_1, X_2, \ldots, X_{100})$ constitute a random sample from a log-normal distribution. Therefore, we equivalently test if the transformed data $Y_i = \log(X_i)$ follow a normal distribution. Figure 2.1 contains a histogram of the log(lifetimes) with a kernel density estimate (Gaussian kernel, $h = .318$) and a normal curve ($\hat{\mu} = 4.84$, $\hat{\sigma} = 1.24$) overlaid. Notice that the histogram and kernel estimate are skewed to the left and the peak is much larger compared to the normal density. Graphically, normality, hence log-normality, appears to be inappropriate.

Under the normal null model, we assume the data come from a $N(\mu, \sigma)$ distribution.

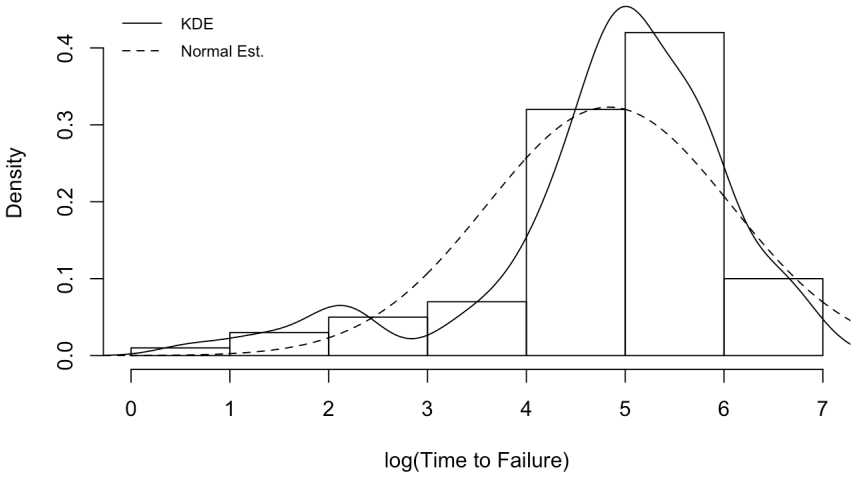

Figure 2.1: Distribution of the $\log$(time to failure) for 100 Kevlar 49 epoxy strands under 80% stress. A kernel density estimate (solid line) and an estimated normal curve (dashed line) are also provided.

Hart and Choi (2016) provide a normal-inverse gamma UIR prior distribution of the form,

$$\pi(\mu, \sigma|\bar{Y}, \gamma) = (2\pi\sigma^2)^{-1/2} \exp\Big[-\frac{1}{2\sigma^2}(\mu - \bar{Y})^2\Big] \frac{2\gamma}{\sqrt{\pi}\sigma^2} \exp\Big[-\frac{\gamma^2}{\sigma^2}\Big], \tag{2.4}$$

where $\gamma = \hat{\sigma}/\sqrt{2}$ for $\hat{\sigma}^2 = \frac{1}{n-m}\sum_{j=m+1}^{n}(Y_j - \bar{Y})^2$ and $\bar{Y} = \frac{1}{n-m}\sum_{j=m+1}^{n} Y_j$. Using the prior in equation (2.4) and normal likelihood for validation data $\mathbf{Y}^V$, the marginal likelihood for the null model is,

$$m(\mathbf{Y}^V|M_0) = \Gamma\Big(\frac{n-m+1}{2}\Big)(n-m+1)^{-(n-m+2)/2}\pi^{-(n-m+1)/2}\hat{\sigma}^{-(n-m)}.$$

As for the alternative model, using the univariate kernel density estimate in (2.1) with Gaussian kernel function and the prior distribution in (2.3), the marginal likelihood is

given by

$$m(\mathbf{Y}^V|M_1) = \int_0^\infty \prod_{j=m+1}^{n} \left[(2\pi)^{-1/2}(mh)^{-1} \sum_{i=1}^{m} \exp\left(-\frac{(Y_j - Y_i)^2}{2h^2} \right) \right] \times \left[\frac{2\beta}{\sqrt{\pi}h^2} \exp\left(-\frac{\beta^2}{h^2} \right) \right] dh. \tag{2.5}$$

Employing the calibration scheme, a $\text{CVWE}_{m,N}$ value is computed for training set sizes $m = \{5, 6, \ldots, 49, 50\}$ using $N = 1,000$ random splits. This curve is plotted in the left panel of Figure 2.2 and is maximized at $m = 30$ with a value equal to 7.241. Next, to ensure the test performs appropriately under the null, 500 random samples are drawn from the estimated null normal model and the $\text{CVWE}_{30,100}$ value for each sample is plotted in the histogram in the right panel of Figure 2.2. All 500 $\text{CVWE}_{30,100}$ values are less than 0 indicating that when $m = 30$, if the observed data were truly normally distributed, the resulting CVWE value should be negative. According to Kass and Raftery (1995), the observed $\text{CVWE}_{30,100}$ value from the Kevlar data of $7.241 > 5$ indicates that there is very strong evidence against the normal model, which implies log-normality of the original times to failure is not appropriate.

The Bayesian nonparametric goodness-of-fit tests of Verdinelli and Wasserman (1998), Berger and Guglielmi (2001), and Tokdar and Martin (2013) were also applied to the Kevlar data. The Verdinelli and Wasserman (1998) method gave the smallest Bayes factor ($\text{BF} = 10$), hence the smallest amount of evidence against the null model. Next, depending on the parameters used in the Berger and Guglielmi (2001) approach, far more evidence in favor of the alternative model is found with Bayes factors between 556 and 1389. Finally, the method by Tokdar and Martin (2013) produced an extremely large Bayes factor ($\text{BF} = 10^5$) against the null model. So for this specific example, the kernel CVBF method finds greater evidence against the null model ($\text{BF} = 1395$) than Verdinelli and Wasserman

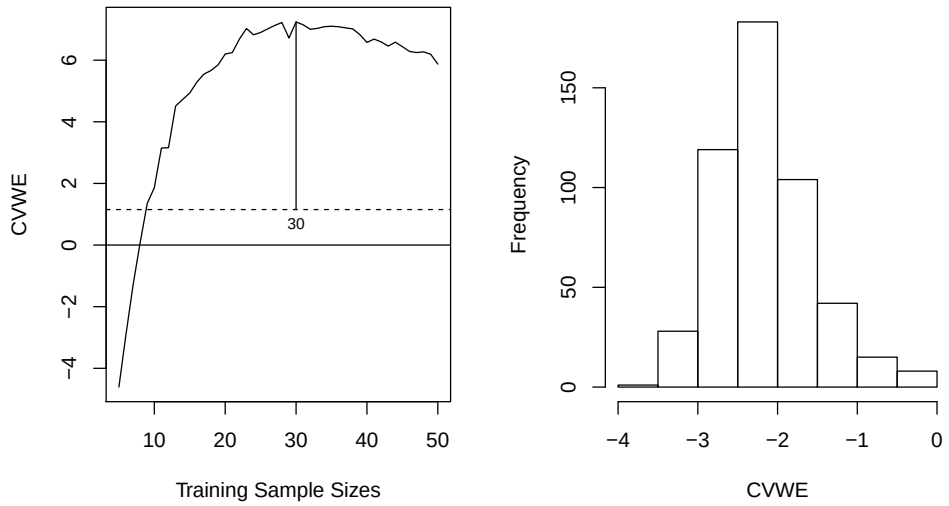

Figure 2.2: *Left Panel*: CVWE values for the observed Kevlar data with $N = 1,000$ random splits at training set sizes $5 \leq m \leq 50$. *Right Panel*: $\text{CVWE}_{30,100}$ values from 500 random samples from the estimated null model.

(1998), a similar amount of evidence as Berger and Guglielmi (2001), and less evidence than Tokdar and Martin (2013).

2.4 Conclusions

This chapter contains sufficient detail to understand the general formulation of the kernel CVBF method in its most basic form before we make any modifications in subsequent chapters. Based on the methodology in Section 2.2, we can see how the CVBF_K method should naturally extend to multivariate data. Also, even though the Kevlar data example in Section 2.3 explores the most common test of normality, the kernel CVBF method can be applied to any parametric null model (see Hart and Choi (2016) for further examples). For our purposes, it makes more sense to include the Kevlar data since it allows for a direct comparison to existing Bayesian nonparametric tests. Even though the kernel

CVBF method falls in between its Bayesian counterparts in terms of the amount of evidence against the normal model, the combination of its performance, intuitiveness, and simplicity make it an attractive alternative nonetheless.

3. MULTIVARIATE KERNEL DENSITY ESTIMATION

In order to carry out the kernel CVBF method for data in more than one dimension we need to better understand the concept of multivariate kernel density estimation. For a detailed description of these methods see the texts by Silverman (1986), Scott (1992), Simonoff (1996), and Wand and Jones (1995).

3.1 Definition

To estimate a d-dimensional multivariate density function f_d for observed data $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ where each $Y_i \in \mathbb{R}^d$, we can use the multivariate kernel density estimate which has the following general form (Wand and Jones, 1995),

$$\hat{f}_d(\mathbf{y}|\mathbf{Y}, \mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2} \sum_{i=1}^{n} K_d\Big(\mathbf{H}^{-1/2}(\mathbf{y} - \mathbf{Y}_i)\Big). \tag{3.1}$$

The first thing to notice in equation (3.1) is that instead of having a scalar smoothing parameter, the multivariate kernel density estimate is indexed by a bandwidth matrix, $\mathbf{H}$. The bandwidth matrix is restricted to the class of symmetric, positive definite matrices, which is an analogous restriction to the scalar bandwidth $h > 0$. Next, the kernel function $K_d(\cdot)$ is typically taken to be a d-variate unimodal probability density function that is symmetric about the origin. There are many possible kernel functions; however, we recommend the d-variate Gaussian kernel,

$$K_d(\mathbf{t}) = (2\pi)^{-d/2} \exp(-\mathbf{t}^T\mathbf{t}/2) = \prod_{l=1}^{d} K_1(t_l),$$

for a variety of reasons. First, the Gaussian kernel is the most common kernel function with noncompact support and satisfies all necessary properties. When using kernel density

estimates to compute a pseudo-likelihood, there is a positive probability that the likelihood function will be 0 for any compact kernel function. Also, the d-variate Gaussian kernel is a product kernel, which means it can be written as the product of univariate Gaussian kernels. This means that we can adapt results from the univariate kernel CVBF method in the necessary derivations for the d-dimensional kernel CVBF method.

3.2 Bandwidth Matrix Classes

In the multivariate kernel density estimate literature, it is common to consider one of three different classes of bandwidth matrices (Wand and Jones, 1995):

- Full (Unconstrained): The class of all symmetric, positive definite bandwidth matrices with $\frac{d(d+1)}{2}$ parameters, denoted as $\mathcal{F}$.
- Diagonal: The class of all diagonal bandwidth matrices with d parameters, $\mathcal{D} = \{\mathbf{H} = \text{diag}(h_1^2, h_2^2, \cdots, h_d^2) : h_l > 0, l = 1, \cdots, d\}$.
- Scalar: The class of all diagonal matrices indexed by a single, scalar bandwidth, $\mathcal{S} = \{\mathbf{H} = h^2\mathbf{I}_d : h > 0\}$.

What are the differences between the three classes? Wand and Jones (1993) give a nice description in terms of the bivariate Gaussian kernel function. Scalar bandwidth matrices restrict the contours of the kernel function to be circular, hence smoothing each coordinate direction of the data by the same amount. For the diagonal bandwidth matrix, the contours are elliptical, but lie parallel to the coordinate axes. Thus, the estimator smooths each coordinate direction by a different amount parallel to the coordinate axes. Finally, the unconstrained bandwidth matrices allow for the contours to be arbitrary ellipses, and thus the estimator smooths in any arbitrary orientation.

3.3 Density Estimation Comparison Across Bandwidth Matrix Classes

There has been some research comparing kernel density estimators based on each of the 3 classes of bandwidth matrices for a variety of densities. Wand and Jones (1993) provide one of the most complete simulation studies with 12 different densities (mixtures of normals) considered. The first notion they point out is that scalar bandwidth matrices should not be used on unscaled multivariate data since the coordinate directions are smoothed by the same amount. Next, when using diagonal bandwidth matrices, the estimator does well for densities where the curvature lies close to the coordinate axes. However, it can be made to do arbitrarily poorly since it does not allow for arbitrary orientations of the data. In order to consider arbitrary orientations of the data while still implementing the kernel estimate based on a diagonal bandwidth matrix, a common technique is to pre-scale and/or pre-smooth the data matrix using the sample covariance matrix (Wand and Jones (1993) and Fukunaga (1990)). The authors point out that this works well for nearly normal densities, but for multimodal densities, these estimators can be made to perform poorly. Therefore, they advise to use an optimal rotation of the data (independent of the covariance matrix) prior to using a diagonal bandwidth matrix as a surrogate for the full matrix. Taking any of these re-scaling approaches would require back transforming the estimator in order to smooth the original data. Their final major conclusion is that for most well behaved densities, considering a diagonal bandwidth matrix is often adequate when compared to the full bandwidth matrix. Of course there are instances where the full bandwidth matrix will be preferred, but in general, a diagonal bandwidth matrix will suffice.

The description from Wand and Jones (1993) compares the efficiencies of the kernel density estimate when using each of the three bandwidth matrix classes for estimation of the true density function. In order to find the best estimate of the unknown density, we need to compute the optimal bandwidth matrix $\mathbf{H}_{\text{opt}}$. There are a wide variety of methods for

finding $\mathbf{H}_{\text{opt}}$, but much like the bandwidth selection problem in univariate kernel density estimation, they typically fall into one of three approaches: standard reference rules, plug-in, and cross-validation (Wand and Jones, 1995).

Standard reference rules are the least sophisticated techniques and require knowledge (or assumption) of the true density function, but they can be used to find $\mathbf{H}_{\text{opt}}$ in any of the three bandwidth matrix classes (Silverman (1986), Scott (1992), Wand and Jones (1993), and Wand and Jones (1995)). Plug-in estimators for $\mathbf{H}_{\text{opt}}$ can also be used for all three classes since at some stage, a standard reference rule is used to estimate a higher order density derivative functional (Wand and Jones (1994) and Duong and Hazelton (2003)). However, compared to using a simple reference rule, using a plug-in approach is often not worth the trouble due to the added complexity of estimating at least one higher order functional. Until recently, more sophisticated cross-validation methods of bandwidth selection that do not require knowledge of the underlying density function were not feasible computationally for unconstrained bandwidth matrices (Sain et al. (1994), Duong and Hazelton (2005), and Zhang et al. (2006)). The Bayesian approach of Zhang et al. (2006) allows us to estimate $\mathbf{H}_{\text{opt}}$ in any of the bandwidth matrix classes using a simple random walk Metropolis-Hastings algorithm.

3.4 Curse of Dimensionality

The term curse of dimensionality takes two different meanings in statistics (Wasserman, 2006). On the one hand, the curse refers to the severe increase in computational burden as the data dimension increases. We see this when we consider the cost/benefit trade-off of choosing the optimal bandwidth matrix within a given bandwidth matrix class and how well we want to estimate the underlying density function in the previous subsection. The scalar bandwidth matrix is the simplest bandwidth matrix to work with as it produces the easiest density estimate to evaluate since as the data dimension increases, the

number of smoothing parameters remains constant at 1. However, due to the inflexibility of the scalar bandwidth matrix, the density estimate is often the least accurate. For a better overall estimate of the true density in most cases, we can consider the diagonal bandwidth matrix class. We do have to pay a small price for a better estimate though. As we increase the data dimension, the number of unknown parameters (d) that we must optimize over to find $\mathbf{H}_{\text{opt}}$ increases linearly. Finally, a full bandwidth matrix would undoubtedly give us the best density estimate in all cases, but now the number of smoothing parameters $(d(d+1)/2)$ scales quadratically with increasing dimension.

On the other hand, Scott (1992) defines the second meaning as the sparsity of data in multiple dimensions. Scott and Thompson (1983) refer to the "empty space phenomenon", which occurs when few observations fall in high-density regions of a multivariate distribution. In order to get an accurate estimate of a multivariate density at a single point, either the smoothing parameter has to be large to include enough observations or the number of observations must be large for the neighborhood to be local. Silverman (1986), Scott and Wand (1991), and Scott (1992) all produce a variety of tables and simulations to show that large data sets are required to estimate the multivariate normal distribution at the zero vector in ten dimensions at the same mean squared error as in one or two dimensions. In all of these references, the general consensus is that kernel density estimation beyond five dimensions is not appropriate in practice.

3.5 Applying Multivariate Kernel Density Estimation to Kernel CVBF

So this discussion leads to two very important questions. First, which one of these three estimation schemes should be used in the multivariate kernel CVBF method? Perhaps we should only consider a full bandwidth matrix class since it gives the best estimate of the underlying density function. However, the computational cost may be too much and in the interest of simplicity, the scalar bandwidth matrix may be preferred. Another possibility

is to take the advice of Wand and Jones (1993) and re-scale the data using the sample covariance matrix before considering a restricted bandwidth matrix class. This way, we can improve our density estimate while still taking advantage of the reduced number of smoothing parameters.

The choice of bandwidth matrix class may also differ depending on the data dimension. When $d = 2$, the computational cost may be inconsequential regardless of bandwidth matrix class, so using $\mathbf{H} \in \mathcal{F}$ may be preferred. As dimension increases, due to the curse of dimensionality, the respective computation times will increase such that eventually $\mathbf{H} \in \mathcal{D}$ and/or $\mathbf{H} \in \mathcal{S}$ become(s) the only feasible option(s).

The second important question is, what are the practical limits on the number of dimensions for which the kernel CVBF method works reasonably well? Of course, the answer to this question depends on the number of observations. Regardless of which bandwidth matrix class we consider, when the data dimension becomes moderately large, accurate estimation of the true density function will become difficult (if not impossible). This could play a pivotal role in determining plausible dimensions for application of the kernel CVBF method. If the kernel model never fits the data well, then we will always favor the null model which makes for a miserable goodness-of-fit test.

Both of these questions will be answered in the next chapter where we consider how to extend the univariate kernel CVBF method of Hart and Choi (2016) to multivariate data.

4. TESTING MULTIVARIATE GOODNESS-OF-FIT USING KERNEL CROSS-VALIDATION BAYES FACTORS

The goal in this chapter is to combine the contents of Chapters 2 and 3 to extend the univariate CVBF$_K$ technique of Hart and Choi (2016) to test goodness-of-fit for data in any dimension. Section 4.1 begins with a description of the overall CVBF$_K$ methodology when applied to multivariate data as slight modifications of the univariate approach must be made. Next, Section 4.2 contains the necessary details for constructing and computing the alternative marginal likelihoods using each of the three bandwidth matrix classes. In order to compare the performance of these three constructions, we carry out simulations in which we test for multivariate normality in Section 4.3. A common theme in this chapter is that we will only consider tests for multivariate normality since the multivariate normal distribution is by far the most common distributional assumption in multivariate analysis and inference. However, keep in mind that the CVBF$_K$ methods can be applied to test any d-dimensional parametric model.

In Section 4.4, we explore the location-scale invariance of the kernel CVBF method and make the necessary modifications to ensure that the resulting conclusions are independent of changes in location and scale. In order to implement the kernel CVBF method in practice, we need to choose the training set size m and the number of random splits N. Section 4.5 describes modifications to the calibration scheme in Subsection 2.2 for finding m as well as a small simulation to explain our recommendation for the choice of N for multivariate data. Arguably the most important property of any model selection technique using Bayes factors is consistency (Definition 1) which will be assessed in Section 4.6 for the scalar bandwidth construction. Also, Section 4.6 includes a description of a *Divide and Conquer* scheme for increasing the computational efficiency of the kernel CVBF method

in large samples without compromising the overall conclusions.

As described in Subsection 1.1.1, there are a few commonly used frequentist tests for goodness-of-fit. Section 4.7 contains a power study for these frequentist tests along with a few kernel CVBF constructions. It is here that we make a final recommendation as to which kernel CVBF construction we recommend in practice after examining their respective performances in terms of power and Type I error rates. However, it will be clear early on in this chapter that the computational burden is far too great for the unconstrained and diagonal bandwidth matrix constructions. One topic that is almost synonymous with multivariate analysis is the curse of dimensionality, which we briefly introduced in Section 3.4. In Section 4.8, we describe how the curse of dimensionality impacts the kernel CVBF methods, in particular its applicability to data beyond moderate dimensions. We also provide possible approaches in which goodness-of-fit can be assessed in higher dimensional data.

Sections 4.2 to 4.8 are all focused on the formulation, properties, and overall performance of the three kernel CVBF constructions. To see how we can assess multivariate goodness-of-fit in practice, Section 4.9 examines testing bivariate normality for Academic Performance Index (API) scores in California schools. In this example we carryout all the calibration steps and illustrate the importance of choosing m appropriately. An interesting application of the kernel CVBF method based on the scalar bandwidth matrix case is in checking the normality assumptions in random effects models. There are some simple modifications to the method that must be made which will be described in Section 4.10. Then, using gene expression data from five rats, we will apply the kernel CVBF method to check the assumptions while also implementing some of the dimension reduction and *Divide and Conquer* techniques described in this chapter. Lastly, an overall summary of this chapter is given in Section 4.11.

4.1 Multivariate Kernel CVBF Methodology

The overall setup of the multivariate kernel CVBF method is very similar to the univariate methodology in Section 2.2. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, $X_i \in \mathbb{R}^d$, comprise a random sample from an unknown d-variate probability density function. The hypotheses we want to test are given by

$$H_0 : \mathbf{X}^V \sim M_0 = \{f_d(\cdot|\theta) : \theta \in \Theta\}$$
$$H_1 : \mathbf{X}^V \sim M_1 = \{\hat{f}_d(\cdot|\mathbf{X}^T, \mathbf{H}) : \mathbf{H} \in \mathcal{S}, \mathcal{D}, \text{or } \mathcal{F}\}.$$

The null model is based on the parametric density function of interest, $f_d(\cdot|\theta)$, and the alternative model requires a family of d-variate kernel density estimates indexed by a $d \times d$ bandwidth matrix from one of the three bandwidth matrix classes.

The original data must again be randomly split into a training set, $\mathbf{X}^T = (X_1, X_2, \ldots, X_m)$, and a validation set, $\mathbf{X}^V = (X_{m+1}, X_{m+2}, \ldots, X_n)$. For a single random split, the Bayes factor in favor of the alternative model can be written as,

$$\text{BF}_m = \frac{\int_{\mathcal{A}} \prod_{j=m+1}^{n} \hat{f}_d(X_j|\mathbf{X}^T, \mathbf{H}) p(\mathbf{H}) d\mathbf{H}}{\int_{\Theta} \prod_{j=m+1}^{n} f_d(X_j|\theta)\pi(\theta) d\theta}. \tag{4.1}$$

Depending on the parametric model being tested, the null marginal likelihood in the denominator of (4.1) may be analytically tractable. In fact, for the normal distribution, we will see that a closed form does exist for the null marginal likelihood when using a common UIR prior distribution for $\pi(\theta)$. The alternative marginal likelihood based on the kernel density estimate is far more complicated, however. Not only does the prior distribution, $p(\mathbf{H})$, change depending on the bandwidth matrix class $\mathcal{A}$, but the bounds of integration (hence the dimension of the integral) change as well. These differences motivate the need to have three different CVBF_K constructions, one for each bandwidth matrix class.

For the most part, the remaining steps of the multivariate kernel CVBF approach directly carry over from the univariate case. The optimal choice for the training set size m is still an open theoretical question, but we can use a modified version of calibration (see Subsection 4.5.1 for the modifications) to make an appropriate choice in practice. Naturally, we expect both the optimal and practical choices of m to be larger proportions of n in the d-dimensional case because more observations are required to adequately estimate the underlying density. Regarding the number of random data splits N to use in practice, typically $30 \leq N \leq 50$ will be more than sufficient. We will provide the justification for this choice in Subsection 4.5.2. One of the main differences between the univariate and d-variate kernel CVBF approaches is how we compute the overall CVWE value for a given data set. For $k = 1, 2, \ldots, N$ random splits of the data matrix into $\mathbf{X}_k^T$ and $\mathbf{X}_k^V$, we compute the weights of evidence $\log(\text{BF}_{m,1}), \ldots, \log(\text{BF}_{m,N})$ and instead of taking the arithmetic mean of the respective weights of evidence, when $d > 1$ we prefer to use

$$\text{CVWE}_{m,N} = \text{median}\Big(\log(\text{BF}_{m,1}), \ldots, \log(\text{BF}_{m,N}) \Big).$$

For n large, the mean and median are comparable for any appropriate choice of N. However, for smaller values of n and m, there are often outlying weights of evidence that cause the mean and median to give contradictory results. Lastly, to determine the strength of the evidence in favor of either the null or alternative models, we continue to use the scale from Kass and Raftery (1995) given in Table 1.2.

4.2 Construction and Computation of the Alternative Marginal Likelihood

In order to carry out the multivariate kernel CVBF method for a given bandwidth matrix class, we require the likelihood function based on the d-dimensional kernel density estimate and the prior distribution $p(\mathbf{H})$. In this section, we provide the form of the likelihood function and the derivation of the prior distribution for each class. Then, we suggest

approaches for numerically approximating the marginal likelihoods. For simplicity, we will denote the $\text{CVBF}_{m,N}$ values constructed using the scalar, diagonal, and unconstrained bandwidth matrix classes as $\text{CVBF}_K(\mathcal{S})$, $\text{CVBF}_K(\mathcal{D})$, and $\text{CVBF}_K(\mathcal{F})$, respectively (similar notation extends to the weights of evidence CVWE).

4.2.1 Scalar Bandwidth Matrix Class : $\text{CVBF}_K(\mathcal{S})$

Under the scalar bandwidth class, the bandwidth matrix used in multivariate kernel density estimation takes the form $\mathbf{H} = h^2\mathbf{I}_d$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix. Thus, the likelihood function based on the multivariate kernel density estimate with Gaussian kernel function reduces to

$$L(\mathbf{X}^V|h, \mathbf{X}^T) = \prod_{j=m+1}^{n} m^{-1}(2\pi h^2)^{-d/2} \sum_{i=1}^{m} \exp\Big(- \frac{[X_j - X_i]^T[X_j - X_i]}{2h^2}\Big), \qquad (4.2)$$

since $\mathbf{H}^{-1} = h^{-2}\mathbf{I}_d$, and $|\mathbf{H}| = h^{2d}$.

We want to formulate the prior distribution $p(h)$ in such a way that it is centered at the validation data and has as much information as a single observation. Let $\mathbf{w}$ represent the d-dimensional vector where each w_l ($l = 1, 2, \ldots, d$) is the median of the l-th column in the validation data. Consider taking $p(h) \propto \hat{f}_d(\mathbf{w}|\mathbf{X}^T, h)$, an evaluation of the likelihood for a single observation at the median. For $d \geq 2$, $\hat{f}_d(\mathbf{w}|\mathbf{X}^T, h)$ is easily integrable when using the Gaussian kernel function. Let $\gamma_i = .5[\mathbf{w} - X_i]^T[\mathbf{w} - X_i]$, then

$$\int_0^\infty m^{-1}(2\pi h^2)^{-d/2} \sum_{i=1}^{m} \exp\Big(- \gamma_i h^{-2}\Big) dh = \frac{1}{2} m^{-1}(2\pi)^{-d/2}\Gamma\Big(\frac{d-1}{2}\Big) \sum_{i=1}^{m} \gamma_i^{-(d-1)/2}.$$

Therefore, $p(h)$ is a proper prior distribution given by

$$p(h|\gamma) = \frac{2\sum_{i=1}^{m} \exp\Big(- \gamma_i/h^2\Big)}{\Gamma\Big(\frac{d-1}{2}\Big) h^d \sum_{i=1}^{m} \gamma_i^{-(d-1)/2}}. \qquad (4.3)$$

Figure 4.1 displays $p(h|\gamma)$ for dimensions $d = 2, 3, 5, 7, 10$ for d-dimensional standard normal data ($n = 1,500$ and $m = 500$). Even though we use a different approach to derive the prior distribution than in the univariate Hart and Choi (2016) method, for $d > 1$ the prior distribution in (4.3) is non-local near 0. In fact, as dimension increases, the neighborhood near 0 where the prior takes very small values gets larger. This aligns with the notion that the optimal smoothing parameter gets larger as dimension increases.

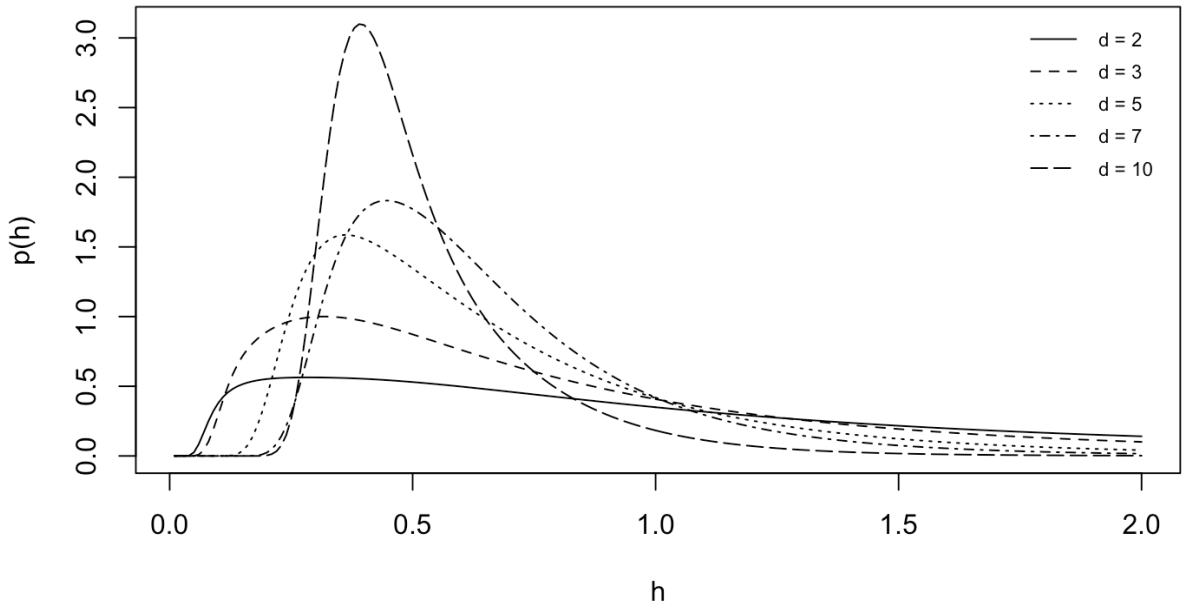

Figure 4.1: Shape of the d-dimensional prior distribution for the scalar bandwidth matrix class.

4.2.2 Diagonal Bandwidth Matrix Class : CVBF$_K(\mathcal{D})$

The bandwidth matrices in the diagonal class have the form, $\mathbf{H} = \text{diag}(h_1^2, h_2^2, \ldots, h_d^2)$ which means the likelihood function with d unknown bandwidth parameters reduces to

$$L(\mathbf{X}^V|h_1, \ldots, h_d, \mathbf{X}^T) = m^{-(n-m)} \prod_{j=m+1}^{n} \sum_{i=1}^{m} \prod_{l=1}^{d} (2\pi h_l^2)^{-1/2} \exp\Big(-\frac{1}{2h_l^2}(X_{jl} - X_{il})^2 \Big), \tag{4.4}$$

since $\mathbf{H}^{-1} = \text{diag}(h_1^{-2}, h_2^{-2}, \ldots, h_d^{-2})$ and $|\mathbf{H}| = \prod_{l=1}^{d} h_l^2$.

To derive the prior distribution for the diagonal bandwidth matrix class, we must consider a different approach to the one we used in the scalar case since taking $p(h_1, \ldots, h_d) \propto \hat{f}_d(\mathbf{w}|\mathbf{X}^T, h_1, \ldots, h_d)$ results in an improper prior distribution. However, finding a proper prior is still a rather straightforward task thanks to the Gaussian kernel being a product kernel and the bandwidth matrix being diagonal. Since the kernel density estimator using a diagonal bandwidth matrix smooths each coordinate independently, we can assume that $p(h_1, \ldots, h_d) = \prod_{l=1}^{d} p(h_l)$. Therefore, $p(h_1, \ldots, h_d)$ will be a proper prior distribution when each $p(h_l)$ is a proper prior distribution. We can apply the IBF approach from Hart and Choi (2016) for finding the univariate prior, by first considering the natural scale invariant improper prior distribution for each smoothing parameter, i.e., $p(h_1, \ldots, h_d) \propto \prod_{l=1}^{d} h_l^{-1}$. Multiplying this improper prior distribution by the form of the likelihood in (4.4) for two random observations from the data (X_1 and X_2), we see that

$$p(h_1, \ldots, h_d) \propto \prod_{l=1}^{d} h_l^{-2} \exp\Big(-\frac{1}{2h_l^2}(X_{jl} - X_{il})^2 \Big). \tag{4.5}$$

The integration of (4.5) is made easier by the fact that we have a product kernel and independent smoothing parameters. In fact, we can integrate over each h_l separately and notice that the resulting prior distributions $p(h_l)$'s have the same form as (2.3) with $\beta_l =$

$\text{IQR}(\mathbf{X}_l^V)/1.35$. Thus, the prior distribution we use in the diagonal bandwidth matrix case is given by

$$p(h_1, \ldots, h_d|\beta) = (4\pi^{-1})^{d/2} \exp\left(-\sum_{l=1}^{d} \frac{\beta_l^2}{h_l^2} \right) \prod_{l=1}^{d} \left(\frac{\beta_l}{h_l^2} \right). \tag{4.6}$$

It is important to point out that the prior distribution in (4.6) is slightly more informative than we would prefer. However, compared to other unit-information priors, this prior is far more stable in practice and for the large sample sizes considered in multivariate analyses, the effect this prior has on the value of the marginal likelihood is negligible. Also, while we do not include graphical displays of the prior distribution in (4.6), the non-local property near the origin is maintained since the d-dimensional prior is the product of univariate non-local priors.

4.2.3 Unconstrained Bandwidth Matrix Class : $\text{CVBF}_K(\mathcal{F})$

Unlike the more restrictive scalar and diagonal bandwidth matrix classes, the likelihood function in the unconstrained bandwidth matrix class does not have a simpler form. Using the general d-variate kernel density estimate, the likelihood function using any symmetric, positive definite bandwidth matrix is given by

$$L(\mathbf{X}^V|\mathbf{H}, \mathbf{X}^T) = \prod_{j=m+1}^{n} [(2\pi)^d m^2 |\mathbf{H}|]^{-1/2} \sum_{i=1}^{m} \exp\Big(-\frac{1}{2}(X_j - X_i)^T \mathbf{H}^{-1}(X_j - X_i) \Big). \tag{4.7}$$

To construct the prior distribution, we again take an IBF approach beginning with an improper prior, namely $p(\mathbf{H}) \propto |\mathbf{H}|^{-d}$. Using the minimum sample size of two observa-

tions, multiplication of the improper prior and the likelihood has the form,

$$\begin{aligned} p(\mathbf{H}) &\propto |\mathbf{H}|^{-d-\frac{1}{2}} \exp\Big(-\frac{1}{2}[X_1 - X_2]^T \mathbf{H}^{-1}[X_1 - X_2]\Big) \\ &= |\mathbf{H}|^{-\frac{2d+1}{2}} \exp\Big(-\frac{1}{2} tr(2\hat{\mathbf{\Sigma}}_V \mathbf{H}^{-1})\Big). \end{aligned} \tag{4.8}$$

Notice that we have used the *trace* operator on the quadratic form, $[X_1 - X_2]^T \mathbf{H}^{-1}[X_1 - X_2]$ and substituted $\hat{\mathbf{\Sigma}}_V$ for $[X_1 - X_2][X_1 - X_2]^T$ since $E\Big([X_1 - X_2][X_1 - X_2]^T\Big) = 2\mathbf{\Sigma}$. Now, the prior distribution in (4.8) is proportional to an Inverse-Wishart$(\nu, \mathbf{\Psi})$ kernel with parameters $\nu = d$ and $\mathbf{\Psi} = 2\hat{\mathbf{\Sigma}}_V$. Therefore, the resulting proper prior distribution is

$$p(\mathbf{H}|\hat{\mathbf{\Sigma}}_V) = \frac{|\hat{\mathbf{\Sigma}}_V|^{d/2}}{\Gamma_d\Big(\frac{d}{2}\Big)} |\mathbf{H}|^{-d-\frac{1}{2}} \exp\Big(-\text{tr}(\hat{\mathbf{\Sigma}}_V \mathbf{H}^{-1})\Big), \tag{4.9}$$

where $\Gamma_d(a) = \pi^{(d-1)/2} \prod_{l=1}^{d} \Gamma\Big(a + \frac{1-l}{2}\Big)$ is the multivariate gamma function.

Notice that we do not begin with the typical Jeffreys' prior $p(\mathbf{H}) \propto |\mathbf{H}|^{-(d+2)/2}$ as our initial improper prior. This is due to the restriction on the degrees of freedom $\nu > d - 1$ in an Inverse-Wishart distribution. This implies that the degrees of freedom must increase with dimension. However, if we were to use Jeffreys' prior distribution in the IBF approach, $\nu = 2$. Thus, the degrees of freedom are constant and our resulting prior would not be a valid Inverse-Wishart distribution when $d > 2$. Therefore, we opt to begin with $p(\mathbf{H}) \propto |\mathbf{H}|^{-d}$ instead.

4.2.4 Numerical Approximation of the Alternative Marginal Likelihood

For all three bandwidth matrix classes, the alternative marginal likelihood is analytically intractable. One common approach in Bayesian analyses is to use Laplace's method to approximate these integrals (Ruli et al., 2016). The multivariate Laplace approximation

is given by

$$\int_{\mathbb{R}^p} \exp(-r(\mathbf{h}))d\mathbf{h} \approx (2\pi)^{d/2}|\hat{\mathbf{V}}|^{-1/2}\exp(-r(\hat{\mathbf{h}})) \tag{4.10}$$

where

- p is the dimension of $\mathbf{h}$, the vector of distinct smoothing parameters.
- $r(\mathbf{h})$ is a smooth and concave function.
- $\hat{\mathbf{h}}$ is the unique minimum of $r(\mathbf{h})$.
- $\hat{\mathbf{V}} = \frac{\partial^2 r(\mathbf{h})}{\partial \mathbf{h} \partial \mathbf{h}^T}$ is the Hessian matrix evaluated at $\hat{\mathbf{h}}$.

In our case, we let $-r(\mathbf{h}) = \log(p(\mathbf{h})L(\mathbf{X}^V|\mathbf{X}^T, \mathbf{h}))$. Experience shows that the kernel likelihood function is bell-shaped, which is the main requirement for the Laplace approximation to be applicable (for more on the required conditions for the appropriateness of the Laplace approximation, see de Bruijn (1961)). All we need to do in order to apply the Laplace approximation to the marginal likelihood is minimize $r(\mathbf{h})$ and find its Hessian matrix, which can often be well-approximated numerically.

In practice, the marginal likelihood for the scalar, diagonal, and unconstrained ($d = 2$) bandwidth matrix classes can be approximated very accurately using a one-, d-, and three-dimensional Laplace approximation, respectively. We can only use a Laplace approximation when $d = 2$ in the full bandwidth matrix class because we must constrain the integral to the class of symmetric, positive definite matrices. In two dimensions, we can write the marginal likelihood as a three-dimensional integral using the constraint, $|h_3| \leq h_1 h_2$, where $\mathbf{H} = \left[\begin{smallmatrix} h_1^2 & h_3 \\ h_3 & h_2^2 \end{smallmatrix}\right]$. For $d > 2$, we must compute the marginal likelihood using some of the approximation techniques found in Evans and Swartz (1995). Essentially, this computation is a typical Bayesian problem which we are trying to avoid in the interest of simplicity. Given the posterior distribution, the integration is difficult due to the large

number of parameters $(d(d+1)/2)$, the complexity of integrating over symmetric positive definite matrices, and the uncertainty of how to carry out the integration efficiently. By using methods like importance sampling or an MCMC approach such as Gibbs sampling or a Metropolis-Hastings algorithm, we require drawing tens of thousands of matrices from an appropriate proposal density. It is pretty easy to see that this will be extremely costly compared to the Laplace approximation in the scalar and diagonal bandwidth matrix cases and should be avoided when possible.

4.3 Testing Multivariate Normality Simulation

In this section we carry out two simulations for testing multivariate normality. The first simulation considers testing bivariate normality using the same distributions as in Wand and Jones (1993), which allows us to compare the relative performance of the three kernel CVBF constructions. The second simulation looks at testing four-dimensional normality for a smaller number of standard distributions. Before we address the simulations, we first derive the null marginal likelihood for the d-variate normal distribution.

4.3.1 Derivation of the Null Marginal Likelihood

In order for the marginal likelihood under the normal model to have a closed form we parameterize the multivariate normal model in terms of the precision matrix $\mathbf{\Psi} = \mathbf{\Sigma}^{-1}$. The likelihood function for the validation data is given by

$$L(\mathbf{X}^V|\mu, \mathbf{\Psi}) = (2\pi)^{-(n-m)/2}|\mathbf{\Psi}|^{(n-m)/2} \exp\Big(-\frac{1}{2} \sum_{j=m+1}^{n} [X_j - \mu]^T \mathbf{\Psi} [X_j - \mu] \Big). \quad (4.11)$$

Under this parameterization, Hoff (2009) provides a UIR prior distribution for $\mu, \mathbf{\Psi}$. Take $\mu|\mathbf{\Psi} \sim N_d(\bar{X}, \mathbf{\Psi}^{-1})$ and $\mathbf{\Psi} \sim \text{Wishart}(d+1, \hat{\mathbf{\Sigma}}^{-1})$, where $\bar{X}$ is the sample mean vector and $\hat{\mathbf{\Sigma}} = (n-m)^{-1} \sum_{j=1}^{n-m} [X_j - \bar{X}][X_j - \bar{X}]^T$ from the validation data. This prior

distribution is as follows:

$$\pi(\mu, \mathbf{\Psi}) = \frac{|\mathbf{\Psi}|^{(n-m)/2} \exp\Big(-\frac{1}{2}[\bar{X} - \mu]^T \mathbf{\Psi}[\bar{X} - \mu]\Big) \exp\Big(-\frac{1}{2} tr(\hat{\mathbf{\Sigma}}\mathbf{\Psi})\Big)}{(2\pi)^{d/2}\Big[2^{(d(d+1)/2)}\Gamma_d\Big(\frac{d+1}{2}\Big)|\hat{\mathbf{\Sigma}}|^{-(d+1)/2}\Big]}. \tag{4.12}$$

Multiplying the prior distribution in (4.12) by the normal likelihood in (4.11) and integrating with respect to μ and $\mathbf{\Psi}$, the marginal likelihood M_0 can be written as

$$M_0 = \pi^{-\frac{d(n-m)}{2}}\left[\frac{\Gamma_d\Big(\frac{n-m+d+1}{2}\Big)}{\Gamma_d\Big(\frac{d+1}{2}\Big)}\right](n-m+1)^{-d(\frac{n-m+d+2}{2})}|\hat{\mathbf{\Sigma}}|^{-(n-m)/2}. \tag{4.13}$$

Now, we can compute the Bayes factor in (4.1) for each of the three kernel CVBF constructions when testing multivariate normality.

4.3.2 Testing Bivariate Normality Simulation

In the following simulation for testing bivariate normality, we use the twelve different mixtures of normal distributions from Wand and Jones (1993) listed below. These distributions cover a wide variety of models with the null model being true for the first two distributions and the alternative model being true for the remaining ten. Note that each component normal distribution is written according to the convention $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

1. Uncorrelated Normal: $N(0, 0, \frac{1}{4}, 1, 0)$
2. Correlated Normal: $N(0, 0, 1, 1, \frac{7}{10})$
3. Skewed: $\frac{1}{5}N(0, 0, 1, 1, 0) + \frac{1}{5}N(\frac{1}{2}, \frac{1}{2}, \frac{4}{9}, \frac{4}{9}, 0) + \frac{3}{5}N(\frac{13}{12}, \frac{13}{12}, \frac{25}{81}, \frac{25}{81}, 0)$
4. Kurtotic: $\frac{2}{3}N(0, 0, 1, 4, \frac{1}{2}) + \frac{1}{3}N(0, 0, 4, 1, -\frac{1}{2})$
5. Bimodal I: $\frac{1}{2}N(-1, 0, \frac{4}{9}, \frac{4}{9}, 0) + \frac{1}{2}N(1, 0, \frac{4}{9}, \frac{4}{9}, 0)$
6. Bimodal II: $\frac{1}{2}N(-\frac{3}{2}, 0, \frac{1}{16}, 1, 0) + \frac{1}{2}N(\frac{3}{2}, 0, \frac{1}{16}, 1, 0)$

7. Bimodal III: $\frac{1}{2}N(-1,1,1,1,\frac{3}{5}) + \frac{1}{2}N(1,-1,1,1,\frac{3}{5})$

8. Bimodal IV: $\frac{1}{2}N(1,-1,\frac{7}{9},\frac{7}{9},0) + \frac{1}{2}N(-1,1,\frac{7}{9},\frac{7}{9},\frac{7}{10})$

9. Trimodal I: $\frac{9}{20}N(-\frac{6}{5},\frac{6}{5},\frac{4}{5},\frac{4}{5},\frac{7}{10}) + \frac{9}{20}N(\frac{6}{5},-\frac{6}{5},\frac{4}{5},\frac{4}{5},-\frac{1}{4}) + \frac{1}{10}N(0,0,\frac{1}{5},\frac{1}{5},\frac{1}{16})$

10. Trimodal II: $\frac{1}{3}N(-\frac{6}{5},0,1,1,\frac{7}{10}) + \frac{1}{3}N(\frac{6}{5},0,1,1,\frac{7}{10}) + \frac{1}{3}N(0,0,1,1,-\frac{7}{10})$

11. Trimodal III: $\frac{3}{7}N(-1,0,\frac{9}{25},\frac{49}{100},\frac{3}{10})+\frac{3}{7}N(1,\frac{2\sqrt{3}}{3},\frac{9}{25},\frac{49}{100},0)+\frac{1}{7}N(1,-\frac{2\sqrt{3}}{3},\frac{9}{25},\frac{49}{100},0)$

12. Quadrimodal: $\frac{1}{8}N(-1,1,1,1,\frac{2}{5}) + \frac{3}{8}N(-1,-1,1,1,\frac{3}{5}) + \frac{1}{8}N(1,-1,1,1,-\frac{7}{10}) + \frac{3}{8}N(1,1,1,1,-\frac{1}{2})$

For each of the twelve distributions, the three kernel CVBF methods will be applied using the following simulation parameters:

- Sample Size: $n = 500$
- Independent Random Samples: 100
- Random Data Splits: $N = 30$
- Training Set Size: $m = 50, 100, 150, 200, 250$

The twelve figures in Appendix A contain the simulation results for each distribution considered. The first panel of each plot displays a contour plot of the true bivariate density function based on the two-dimensional kernel density estimate with each coordinate smoothed using the same normal reference bandwidth. The second panel provides results for each of the three kernel CVBF methods applied to the same 500 random samples. The solid, dashed, and dotted curves represent the median CVWE values from the CVBF$_K(\mathcal{S})$, CVBF$_K(\mathcal{D})$, and CVBF$_K(\mathcal{F})$ methods, respectively. Also, the vertical lines correspond interquartile range of CVWE values with endpoints at the first and third quartiles.

When the true density is a normal distribution, all three curves increase monotonically to 0 as the training set size increases, without reaching 0. For the uncorrelated normal model, only when $m = 50$ does the scalar method find much stronger evidence in favor of the normal model compared to the other two constructions. At the remaining four training set sizes, the three kernel CVBF methods produce comparable results. When the two coordinates are correlated, the $\text{CVBF}_K(\mathcal{D})$ method produces the lowest CVWE values at all training set sizes. So when the null hypothesis is true for testing normality, it appears that $\text{CVBF}_K(\mathcal{S})$ is preferred when the coordinates are uncorrelated and $\text{CVBF}_K(\mathcal{D})$ when correlation is present. That being said, the CVWE values for all three constructions correctly find strong evidence in favor of the normal model based on the Kass and Raftery (1995) criterion, CVWE $< -\log(20)$, at any training set size $m \in [50, 250]$.

For the remaining ten distributions in which the alternative model is true, the three curves tend to have the following relationship: $\text{CVWE}_K(\mathcal{D}) \leq \text{CVWE}_K(\mathcal{F}) \leq \text{CVWE}_K(\mathcal{S})$. This relationship between the three CVBF constructions holds for all distributions except the bimodal II density in Figure A.6 if we consider the training set size that maximizes each of the three CVWE curves. However, in this instance, the training set size is $m = 50$, which in practice is too small to adequately estimate the true bivariate density. For suitable choices of the training set size, the $\text{CVBF}_K(\mathcal{S})$ method finds the strongest evidence against the normal model in the ten alternative models considered.

Considering the results from the twelve bivariate distributions as a whole, the $\text{CVBF}_K(\mathcal{S})$ method generally provides the strongest conclusions in favor of the correct hypothesis when testing bivariate normality. That being said, the $\text{CVBF}_K(\mathcal{D})$ and $\text{CVBF}_K(\mathcal{F})$ approaches also perform very well in that they too favor the true model in all twelve cases. The real distinction between the respective performances of these three constructions is the drastic difference in computation time. In order to compute the kernel CVBF value for a single data set of $n = 500$ bivariate normal observations, the $\text{CVBF}_K(\mathcal{D})$ and $\text{CVBF}_K(\mathcal{F})$

methods take about 4 and 400 times longer, respectively, compared to the CVBF$_K(\mathcal{S})$ method. This is extremely intriguing because we often have to pay a penalty in the interest of a simpler method. Yet here, the simplest and most intuitive kernel CVBF approach performs the best *and* is the fastest to compute.

Overall, this simulation for testing bivariate normality shows that the CVBF$_K(\mathcal{F})$ method is not worth pursuing further because of its computational inefficiency and mediocre performance. Our goal is to find a simple and intuitive Bayesian method for testing goodness-of-fit and certainly we have two approaches, namely CVBF$_K(\mathcal{S})$ and CVBF$_K(\mathcal{D})$, that better suit this goal compared to CVBF$_K(\mathcal{F})$. Even though the CVBF$_K(\mathcal{D})$ method is slightly more computationally demanding compared to the CVBF$_K(\mathcal{S})$ method, it did outperform its counterparts in certain cases. Therefore, in the simulations and discussion to follow we will still explore both the CVBF$_K(\mathcal{S})$ and CVBF$_K(\mathcal{D})$ approaches.

4.3.3 Testing d-variate Normality Simulation

As we saw for bivariate data, all three CVBF$_K$ constructions perform quite well when testing normality, most notably the CVBF$_K(\mathcal{S})$ method. How well do these constructions perform if we consider more than two dimensions when testing normality?

To answer this question we consider another simulation, but this time we will test four-dimensional normality. In this simulation, we only consider four distributions:

- Standard Normal Distribution: $N(\mu = \mathbf{0}, \mathbf{\Sigma} = \mathbf{I}_4)$,
- Independent Laplace Distribution: each coordinate vector follows a Laplace$(\mu = 0, b = 1)$ distribution,
- t Distribution (df = 3): $t_3(\mu = \mathbf{0}, \mathbf{\Sigma} = \mathbf{I}_4)$, and
- Skew Normal Distribution: $SN(\xi = \mathbf{0}, \mathbf{\Omega} = \mathbf{I}_4, \alpha = \mathbf{10})$.

These four distributions were chosen such that we can not only see performance under the null hypothesis (standard normal) but also under a few alternative models that encompass the most common departures from normality: peakedness (Laplace), heavy tails (t_3), and skewness (skew-normal). A bimodal mixture distribution was also included in the simulation, but, the resulting CVWE values were so large that we merely report this fact rather than providing plots. In total, 100 independent samples of size $n = 2,000$ are drawn from each of the four distributions. Each sample is then randomly split $N = 30$ times for training set sizes $m = 200, 400, 600, 800,$ and 1000. Figure 4.2 displays the resulting $\text{CVWE}_K(\mathcal{S})$ and $\text{CVWE}_K(\mathcal{D})$ values, respectively. We do not consider the $\text{CVBF}_K(\mathcal{F})$ construction beyond two dimensions based on the discussion in the previous subsection.

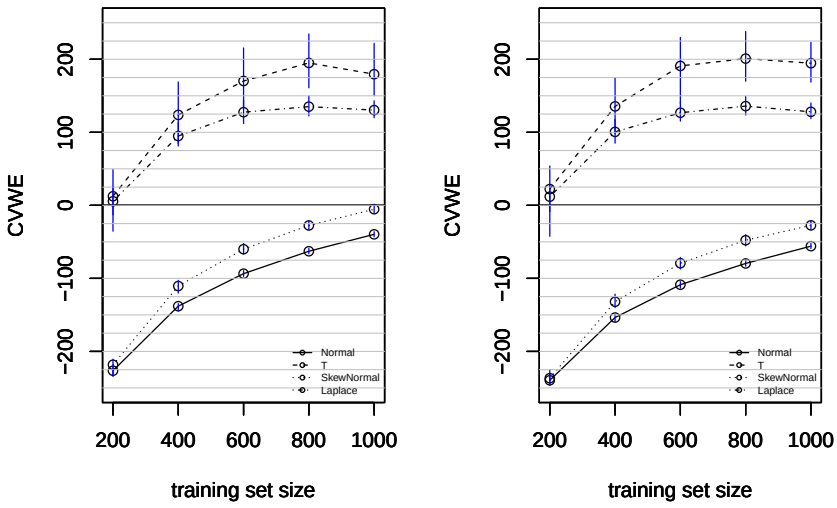

Figure 4.2: Testing 4-D normality using $\text{CVWE}_K(\mathcal{S})$ (*left panel*) and $\text{CVWE}_K(\mathcal{D})$ (*right panel*) for 100 random samples ($n = 2000$) from a standard normal distribution (solid), t_3 distribution (dashed), skew-normal distribution (dotted), and Laplace distribution (dotdashed). Each sample is randomly split $N = 30$ times for training set sizes $m = 200, 400, 600, 800,$ and 1000

The first thing to notice straight away in Figure 4.2 is the overall similarity between the two panels in that the CVWE_K curves only differ slightly across the two constructions. This should not come as a surprise because the coordinate variances are equal in each of the four distributions considered in this simulation. In fact, the normal, t_3, and Laplace distributions all have covariance matrices proportional to the identity matrix. Therefore, the kernel density estimates of the underlying density functions should be similar for both constructions. For these three distributions, the $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ methods have no problem reaching the correct conclusions.

The skew-normal distribution proves to be problematic in that we would incorrectly conclude in favor of normality for all training set sizes in both CVBF_K constructions. We believe the poor performance of both $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ stems from a combination of two factors. First, unlike the other three distributions in this simulation, the covariance matrix of our skew-normal model is not proportional to the identity matrix (the impact of the covariance matrix on the kernel CVBF methods will be explored further in the next section). The true covariance matrix for the skew-normal distribution when $\xi = \mathbf{0}$ is given by $\mathbf{\Sigma} = \mathbf{\Omega} - \mu\mu^t$, where $\mu = \left(\frac{2}{\pi(1+\alpha^T\mathbf{\Omega}\alpha)}\right)^{1/2}\mathbf{\Omega}\alpha$ (Azzalini and Capitanio, 1998). Plugging in our parameters, $\mathbf{\Sigma} \approx \mathbf{I}_4 - .16\mathbf{J}_4$, where $\mathbf{J}_d$ is a $d \times d$ matrix with each element equal to 1. Secondly, with so few observations in the training set, the kernel density estimate cannot adequately detect the skewness. Both of these factors lead to the estimated parametric model serving as a better representation of the skew-normal data compared to the kernel model based on either a scalar or diagonal bandwidth matrix.

To better explain this phenomenon, we compared each of the true alternative models to both the multivariate kernel density estimate and estimated multivariate normal model using Kullback-Leibler discrepancies. Define f to be the true alternative model, $\tilde{f}$ to be the estimated multivariate null model using the maximum likelihood parameter estimates, and $\hat{f}$ to be the kernel density estimate based on the optimal scalar bandwidth matrix chosen

using a standard reference rule. A crude approximation to $\log(\text{BF}_m)$ for a single random split in (4.1) is the log likelihood ratio $\log(\Lambda)$. It can be shown that $\log(\Lambda)$:

$$\begin{aligned}\log\Lambda &= \sum_{j=m+1}^{n}\log(\hat{f}(\mathbf{X}_j)) - \sum_{j=m+1}^{n}\log(\tilde{f}(\mathbf{X}_j)) \\ &\approx (n-m)\Big[\int \log\Big(\frac{\hat{f}(\mathbf{x})}{f(\mathbf{x})}\Big)f(\mathbf{x})d\mathbf{x} - \int \log\Big(\frac{\tilde{f}(\mathbf{x})}{f(\mathbf{x})}\Big)f(\mathbf{x})d\mathbf{x}\Big] \\ &= (n-m)\big[\text{KL}(\tilde{f},f) - \text{KL}(\hat{f},f)\big], \end{aligned} \tag{4.14}$$

where $\text{KL}(g,f)$ denotes the Kullback-Leibler divergence. Expression (4.14) provides further intuition as to what is happening in the kernel CVBF method. When $\log(\Lambda) > 0$, the optimal kernel model is closer to the true density compared to the estimated parametric model and vice versa when $\log(\Lambda) < 0$. If we compute the approximation in (4.14) for the three alternative models, the resulting values should be positive since the alternative hypothesis is true. After 25 independent samples of size $n = 2,000$ from the skew-normal, t_3, and Laplace distributions with training set size $m = 400$, the approximate average log likelihood ratios are -175.47, 32.28, and 44.96, respectively. What we learn from this exploration is that the skew-normal distribution is closer to the estimated normal model than the optimal kernel density estimate in a Kullback-Leibler sense. Therefore, it is not surprising that the CVBF_K methods favor normality over skew-normality.

4.3.4 Simulation Conclusions

Overall, both of these simulations for testing multivariate normality indicate that the univariate CVBF_K method of Hart and Choi (2016) can be extended to test multivariate goodness-of-fit. From the bivariate simulation, we saw that for a variety of distributions all three formulations from Section 4.2 performed similarly well when testing normality. With a moderate sample size of $n = 500$, the $\text{CVBF}_K(\mathcal{S})$ approach performed the best in general. However, the most telling difference between the three constructions is their

respective computation times. The $\text{CVBF}_K(\mathcal{F})$ is certainly not useful in practice since its computation time is immensely longer than the other two approaches without drastically outperforming them.

The simulation results for testing four-dimensional normality were positive, but left us with a few unanswered questions. Both the $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ methods performed very well under the null model for standard normal data. However, under the alternative models, both approaches correctly detected departures from normality in the form of peakedness and heavy tails, but struggled to detect skewness. Due to their similar performances, at this point we do not have a recommendation in favor of either the $\text{CVBF}_K(\mathcal{S})$ or $\text{CVBF}_K(\mathcal{D})$ construction. In subsequent sections we will look at different d-dimensional distributions from the four families considered in Subsection 4.3.3 in order to better compare these two approaches as well as to try and remedy their difficulty in detecting skewness.

4.4 Effect of Location and Scale on Kernel CVBF

Any goodness-of-fit test should be location-scale invariant. In terms of the kernel CVBF method, this means that if we change the location and/or scale of the data, the resulting CVBF values remain the same. When testing univariate normality, Hart and Choi (2016) showed that the Bayes factor in (2.2) is location-scale invariant when using the UIR prior in (2.4). Does this invariance property extend to the scenario of testing multivariate normality?

4.4.1 Location Invariance

Let $\mathbf{X}$ continue to represent the random sample of n observations from the true d-variate density function. By definition, the kernel CVBF method is invariant to changes in location if the $\text{CVBF}_{m,N}$ value computed from $\mathbf{X}$ equals the $\text{CVBF}_{m,N}$ value computed from $Y_i = X_i + \mathbf{c}$, for constant vector $\mathbf{c} \in \mathbb{R}^d$, for the same N random splits. Provided that

we use a UIR prior distribution for the null model, it is easy to see that the kernel CVBF method based on any of the three bandwidth matrix classes is location invariant.

In the derivation of the normal marginal likelihood in Subsection 4.3.1, we see that the only term in (4.13) that depends on the data vectors is $|\hat{\mathbf{\Sigma}}_{\mathbf{X}}|^{-(n-m)/2}$. If we compute the sample covariance matrix from the validation data $\mathbf{Y}^V$, we see that

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{\mathbf{Y}} &= \frac{1}{n-m}\sum_{j=m+1}^{n}[Y_j-\bar{Y}][Y_j-\bar{Y}]^T \\ &= \sum_{j=m+1}^{n}[X_j+\mathbf{c}-\bar{X}-\mathbf{c}][X_j+\mathbf{c}-\bar{X}-\mathbf{c}]^T \\ &= \hat{\mathbf{\Sigma}}_{\mathbf{X}}.\end{aligned}$$

In Subsections 4.2.1-4.2.3, the three kernel marginals in the CVBF constructions depend on the sample covariance matrix and/or the following quantities:

- $[X_j - X_i]^T[X_j - X_i]$,
- $\gamma_i = \frac{1}{2}[\mathbf{w} - X_i]^T[\mathbf{w} - X_i]$, where $\mathbf{w}$ is a vector of column medians of the training data,
- $\beta_l = \text{IQR}(\mathbf{X}_{.l}^V)/1.35$, and
- $[X_j - X_i]^T\mathbf{H}^{-1}[X_j - X_i]$.

Noting that the interquartile range is location invariant and $\text{median}(\mathbf{Y}_{.l}^V) = \text{median}(\mathbf{X}_{.l}^V) + c_l$, all of these terms are invariant to location changes by the same simple argument as for the sample covariance matrix. Therefore, changing the center of the observed data vectors does not effect the resulting kernel CVBF value.

4.4.2 Scale Invariance

In order to show that the kernel CVBF method is scale invariant, we want to show that the CVBF value remains unchanged whether we use the original data $\mathbf{X}$ or scale transformed data $Y_i = \mathbf{A}X_i$, where $\mathbf{A}$ is an invertible $d \times d$ matrix of constants. Unlike the previous subsection where we get a cancellation of the location change in the null and alternative marginal likelihoods, the changes in scale do not always cancel out so nicely.

Beginning again with the null marginal likelihood, the sample covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{Y}} = \mathbf{A}\hat{\mathbf{\Sigma}}_{\mathbf{X}}\mathbf{A}^T$ using standard properties of covariance. This means that

$$|\hat{\mathbf{\Sigma}}_{\mathbf{Y}}|^{-(n-m)/2} = |\mathbf{A}|^{-(n-m)}|\hat{\mathbf{\Sigma}}_{\mathbf{X}}|^{-(n-m)/2}.$$

Therefore, in order for the kernel CVBF method to be scale invariant, the alternative marginal likelihood must contain the factor $|\mathbf{A}|^{-(n-m)}$ that cancels under each of the three bandwidth matrix classes.

For both the scalar and diagonal bandwidth matrix classes, scale invariance only occurs for specific transformations. First consider scale invariance of the $\text{CVBF}_K(\mathcal{D})$ construction. Let $\mathbf{A} = \text{diag}(a_1, \ldots, a_l)$ with $a_l \neq 0$ for $l = 1, 2, \ldots, d$. The alternative marginal likelihood computed from the transformed data $\boldsymbol{Y}$ is as follows:

$$\begin{aligned}
&\int \cdots \int L(\boldsymbol{Y}^V|\boldsymbol{Y}^T, h_1, \ldots h_d)p(h_1, \ldots, h_d)dh_1 \cdots dh_d = \\
&C \int \cdots \int \left[\prod_{j=m+1}^{n} \sum_{i=1}^{m} \prod_{l=1}^{d} h_l^{-1} \exp\Big(- \frac{(a_l X_{jl} - a_l X_{jl})^2}{2h_l^2} \Big) \right] \\
&\qquad \times \exp\Big(- \sum_{l=1}^{d} \frac{a_l^2 \beta_l}{h_l^2} \Big) \prod_{l=1}^{d} \frac{a_l \beta_l}{h_l^2} dh_1 \cdots dh_d,
\end{aligned}$$

where C is the appropriate normalizing constant unaffected by changes in scale. Now, consider the change of variables $b_l = h_l/a_l$ for $l = 1, \ldots, d$. Using the fact that the

required Jacobian is $|\mathbf{A}|^{-1} = \prod_{l=1}^{d} a_l^{-1}$, the marginal likelihood reduces to

$$C|\mathbf{A}|^{-(n-m)} \int \cdots \int \left[\prod_{j=m+1}^{n} \sum_{i=1}^{m} \prod_{l=1}^{d} b_l^{-1} \exp\Big(- \frac{(X_{jl} - X_{jl})^2}{2b_l^2} \Big) \right]$$
$$\times \exp\Big(- \sum_{l=1}^{d} \frac{\beta_l^*}{b_l^2} \Big) |\mathbf{A}| \prod_{l=1}^{d} \frac{\beta_l^*}{b_l^2} |\mathbf{A}|^{-1} db_1 \cdots db_d,$$

where β_l^* is β_l in (4.6) for the data $\boldsymbol{X}$. If we rewrite this integral in terms of the likelihood function in (4.4) and prior distribution p^* in (4.6) for the data $\boldsymbol{X}$, the marginal likelihood for $\boldsymbol{Y}$ equals

$$|\mathbf{A}|^{-(n-m)} \int \cdots \int L(\boldsymbol{X}^V | \boldsymbol{X}^T, b_1, \ldots b_d) p^*(b_1, \ldots, b_d) db_1 \cdots db_d.$$

Therefore, the $\text{CVBF}_K(\mathcal{D})$ method is scale invariant for the transformation $Y_i = \mathbf{A}X_i$ when $\mathbf{A}$ is a diagonal matrix since we get the cancellation of $|\mathbf{A}|^{-(n-m)}$ with the null marginal likelihood.

A similar result holds in the scalar bandwidth matrix case. Consider the same transformation $Y_i = \mathbf{A}X_i$, but let $\mathbf{A} = a\mathbf{I}_d$ for $a \neq 0$. For the likelihood function in (4.2) and prior distribution in (4.3), the alternative marginal likelihood for the transformed data $\boldsymbol{Y}$ reduces to

$$\begin{aligned} \int L(\boldsymbol{Y}^V | \boldsymbol{Y}^T, h) p(h|\gamma_Y) dh &= \int L(\boldsymbol{X}^V | \boldsymbol{X}^T, b) |\mathbf{A}|^{-(n-m)} p(b|\gamma_X) a a^{-1} db \\ &= |\mathbf{A}|^{-(n-m)} \int L(\boldsymbol{X}^V | \boldsymbol{X}^T, b) p(b|\gamma_X) db. \end{aligned}$$

Thus, under this scalar transformation, the $\text{CVBF}_K(\mathcal{S})$ method is scale invariant.

Unlike the $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ constructions, the $\text{CVBF}_K(\mathcal{F})$ method is scale invariant for any invertible constant matrix $\mathbf{A}$. Once again, consider the alternative marginal

likelihood for $\boldsymbol{Y}$ using (4.7) and (4.9). After substituting in $Y_i = \mathbf{A}X_i$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}} = \mathbf{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\mathbf{A}^T$, the marginal likelihood is as follows:

$$C\int \prod_{j=m+1}^{n} |\mathbf{H}|^{-1/2} \sum_{i=1}^{m} \exp\Big(-\frac{1}{2}(X_j - X_i)^T\mathbf{A}^T\mathbf{H}^{-1}\mathbf{A}(X_j - X_i)\Big)$$
$$\times|\mathbf{A}|^d|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}|^{d/2}|\mathbf{H}|^{-d-1/2}\exp\Big(-tr(\mathbf{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\mathbf{A}^T\mathbf{H}^{-1})\Big)d\mathbf{H}.$$

Let $\mathbf{B}^{-1} = \mathbf{A}^T\mathbf{H}^{-1}\mathbf{A}$ such that $\mathbf{H} = \mathbf{A}\mathbf{B}\mathbf{A}^T$. Making the change of variables, the marginal likelihood reduces to

$$C\int |\mathbf{A}|^{-(n-m)} \prod_{j=m+1}^{n} |\mathbf{B}|^{-1/2} \sum_{i=1}^{m} \exp\Big(-\frac{1}{2}(X_j - X_i)^T\mathbf{B}^{-1}(X_j - X_i)\Big)$$
$$\times|\mathbf{A}|^d|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}|^{d/2}|\mathbf{A}|^{-2d-1}|\mathbf{B}|^{-d-1/2}\exp\Big(-tr(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\mathbf{B}^{-1})\Big)|\mathbf{A}|^{d+1}d\mathbf{B}.$$

When making this change of variables, we have implicitly stated that $d\mathbf{H} = |\mathbf{A}|^{d+1}d\mathbf{B}$. In order to determine that $|\mathbf{A}|^{d+1}$ is the appropriate Jacobian, consider making this change of variables to only the prior distribution, which we know must integrate to 1. Therefore, we know that

$$|\mathbf{A}|^{-(d+1)}\mathbf{J}\int \frac{|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}|^{d/2}}{\Gamma_d\Big(\frac{d}{2}\Big)}|\mathbf{B}|^{-d-1/2}\exp\Big(-\text{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\mathbf{B}^{-1})\Big)d\mathbf{B} = 1,$$

where we have made the necessary substitutions and $\mathbf{J}$ is the unknown Jacobian. Clearly, $\mathbf{J} = |\mathbf{A}|^{d+1}$ since the integral of an Inverse-Wishart distribution is 1. Thus, the final form of the marginal likelihood for $\boldsymbol{Y}$ is given by

$$|\mathbf{A}|^{-(n-m)}\int L(\boldsymbol{X}^V|\boldsymbol{X}^T,\mathbf{B})p(\mathbf{B}|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}})d\mathbf{B}.$$

Hence, the CVBF$_K(\mathcal{F})$ method is scale invariant.

Overall, while the kernel CVBF method for testing multivariate normality is location invariant, the $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ methods cannot be made scale invariant. The scalar and diagonal kernel CVBF methods are only scale invariant under specific transformations, namely, the transformation matrix must be a member of each of their respective bandwidth matrix classes. In practice, these restricted transformations are only useful in a handful of scenarios. For instance, the diagonal transformation would be applicable when the coordinates are known to be pairwise independent and the scalar transformation assumes each coordinate has the same variance in addition to pairwise independence. The $\text{CVBF}_K(\mathcal{F})$ method is the only kernel CVBF approach that is both location and scale invariant for general data transformations. This lack of scale invariance across all bandwidth matrix classes is an unfortunate difference from the univariate approach of Hart and Choi (2016) even though we also utilize a UIR prior distribution for the parameters under the null model.

4.4.3 Location-Scale Invariant Version of the $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ Methods

An additional price we have to pay for using a restricted bandwidth matrix in place of the full bandwidth matrix is a lack of scale invariance. This means that the amount of evidence in favor of the null model depends on the underlying covariance matrix of the true density function. For bivariate normal data for instance, the $\text{CVBF}_K(\mathcal{S})$ value decreases as the correlation between the two coordinates tends toward ± 1. Therefore, every d-dimensional normal model with a different covariance matrix will produce a different $\text{CVBF}_K(\mathcal{S})$ value. Similar dependencies may exist for other parametric null models, which is less than ideal from a goodness-of-fit perspective. So what alternative approaches (if any) can we explore that will in effect make the $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ methods location-scale invariant?

Back in Section 3.3, we summarized the recommendations of Wand and Jones (1993)

regarding when to use each bandwidth matrix class for multivariate density estimation. They strongly recommended avoiding the scalar bandwidth matrix unless the data are re-scaled so that it would make more sense to smooth each coordinate by the same amount. Similarly, in most cases, the full bandwidth matrix can be replaced with the diagonal bandwidth matrix provided that the data are re-scaled appropriately. Staying consistent with their recommendations, if we re-scale the data using the sample covariance matrix based on all n data vectors, so that $Y_i = \hat{\mathbf{\Sigma}}_{\mathbf{X}}^{-1/2} X_i$, then $\hat{\mathbf{\Sigma}}_{\mathbf{Y}} = \mathbf{I}_d$. This transformation is especially appealing in the case of testing normality because testing $\mathbf{X} \sim N_d(\mu, \mathbf{\Sigma})$ is equivalent to testing $\mathbf{\Sigma}^{-1/2}\boldsymbol{X} \sim N_d(\mathbf{\Sigma}^{-1/2}\mu, \mathbf{I}_d)$. Also, thinking back to the situations where the scale invariant transformations for CVBF$_K(\mathcal{S})$ or CVBF$_K(\mathcal{D})$ are appropriate, the transformed data $\boldsymbol{Y}$ have identity sample covariance. Now, provided that we re-scale the data to have identity sample covariance prior to applying the kernel CVBF method, location *and* scale invariance holds and it makes perfect sense to use either the CVBF$_K(\mathcal{S})$ or CVBF$_K(\mathcal{D})$ construction.

4.4.4 Simulation Results for CVBF$_K(\mathcal{S})$ on Re-Scaled Observations

How does this transformation approach perform when testing four-dimensional normality? Remember back in Subsection 4.3.3, three of the four distributions already had sample covariance matrices proportional to the identity matrix. Certainly, we expect comparable performance for these distributions, but they will be slightly different due to the constant of proportionality. The real question will be how do the CVBF results differ for the skew-normal distribution after applying the scale transformation. The first simulation in this subsection explores the performance of the CVBF$_K(\mathcal{S})$ method on the scale transformed data from the three alternative distributions in Subsection 4.3.3. The resulting median CVWE values from this simulation (same distributions and parameters as before) are provided in Figure 4.3. We exclude the standard normal data from this simulation since

$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} \approx \mathbf{I}_4$. Thus, we expect the resulting CVWE values to be approximately the same as those in Figure 4.2.

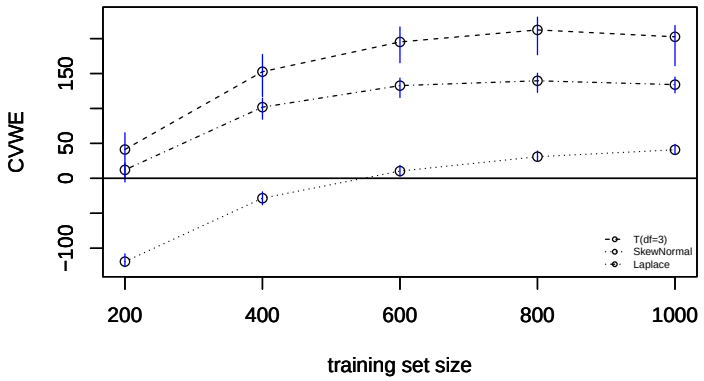

Figure 4.3: Testing 4-D normality using $\text{CVWE}_K(\mathcal{S})$ for re-scaled data from a t_3 distribution (dashed), skew-normal distribution (dotted), and Laplace distribution (dotdashed). In total, 100 independent random samples of size $n = 2000$ are considered for each distribution and the CVWE values are based on $N = 30$ splits and training set sizes $m = 200, 400, 600, 800,$ and 1000.

For the t_3 and Laplace distributions, the CVWE curves in Figure 4.3 are comparable to those in the left panel of Figure 4.2. In fact, by re-scaling the data prior to computing the $\text{CVWE}_K(\mathcal{S})$ values, we see much larger CVWE values for the smaller training set sizes $m = 200, 400$. Regarding the skew-normal model, notice the vast improvement in Figure 4.3. When using the raw observations, the $\text{CVWE}_K(\mathcal{S})$ values in Figure 4.2 were negative for all training set sizes. Now, after transforming the data to have identity sample covariance, provided that $m \geq 600$, we would correctly conclude against normality.

A more logical simulation to explore the performance of this location-scale invariant approach to the original $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ constructions is to consider members

of these four families (t_3, normal, skew-normal, and Laplace) that do not have covariance matrices proportional to identity. Consider the following four trivariate distributions:

- Normal: $\mu = (3.4, 5.5, 3.5)^T$, $\mathbf{\Sigma} = \begin{bmatrix} 5.5 & 2.1 & -.2 \\ 2.1 & 2.0 & .02 \\ -.2 & .02 & 9.9 \end{bmatrix}$,
- Skew-Normal: $\xi = (-14.1, 18.9, 15.5)^T$, $\mathbf{\Omega} = \begin{bmatrix} 5.5 & -3.9 & 1.3 \\ -3.9 & 5.1 & -1.6 \\ 1.3 & -1.6 & 2.1 \end{bmatrix}$, $\alpha = (15.9, 7.1, -6.0)^T$,
- t_3: $\mu = (0, 0, 0)^T$, $\mathbf{\Sigma} = \begin{bmatrix} 7.0 & -2.0 & 3.1 \\ -2.0 & 4.4 & 0.5 \\ 3.1 & 0.5 & 3.5 \end{bmatrix}$, and
- Laplace: $\mu = (-8.2, -6.6, 5.3)^T$, $\lambda = (1.4, 0.8, 12.7)^T$.

Figure 4.4 contains the median CVWE values from the scale invariant CVBF$_K(\mathcal{S})$ approach and the original CVBF$_K(\mathcal{S})$ and CVBF$_K(\mathcal{D})$ methods. From each distribution, we drew 96 independent random samples of $n = 1,000$ observations.

In the top right panel of Figure 4.4, all three kernel CVBF methods perform similarly well when the null hypothesis is true. The real difference exists when the alternative hypothesis is true. First, when the sample covariance matrix is not proportional to the identity matrix, the standard CVBF$_K(\mathcal{S})$ approach performs very poorly. In fact, for each of the alternative models, we would strongly favor normality. Next, the CVBF$_K(\mathcal{D})$ approach correctly favors the kernel model for the t_3 (bottom left panel) and Laplace distributions (bottom right panel), but not the skew-normal model (top right panel). The only kernel CVBF method that reaches the appropriate conclusion in the four distributions for suitable training set sizes is the CVBF$_K(\mathcal{S})$ method computed on the transformed data. In subsequent sections, "the scaled CVBF$_K(\mathcal{S})$ method" will refer to the application of the original CVBF$_K(\mathcal{S})$ method to re-scaled data.

4.4.5 Summary

Overall, this section illustrates a few very important aspects of applying the kernel CVBF method to distributions with varying centers and scales. The original constructions

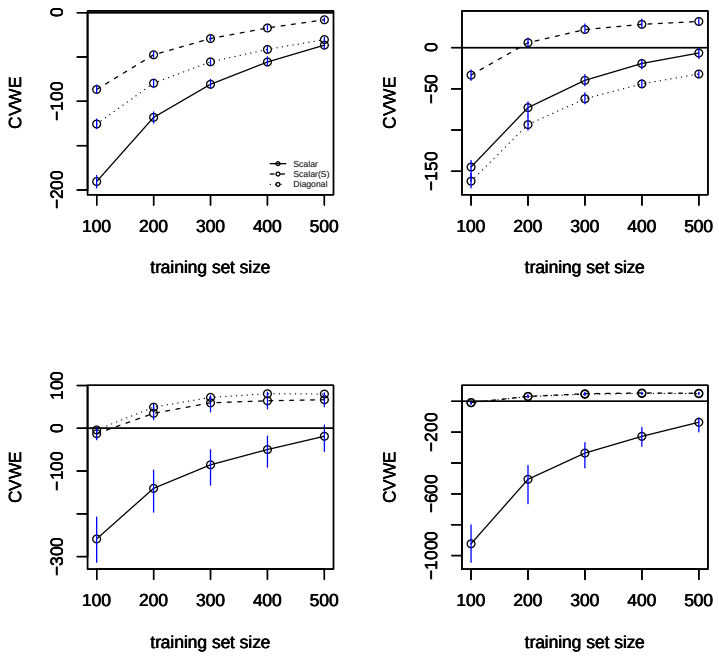

Figure 4.4: Testing 3-D normality using $\text{CVWE}_K(\mathcal{S})$ on the original data (solid curves) and re-scaled data (dashed curves) as well as $\text{CVWE}_K(\mathcal{D})$ on the original data (dotted curves). In total, 96 random samples of size $n = 1,000$ were drawn from the normal (*top left panel*), skew-normal (*top right panel*), t_3 (*bottom left panel*), and Laplace (*bottom right panel*) distributions.

of Section 4.2 are location invariant, but only the $\text{CVBF}_K(\mathcal{F})$ method is scale invariant. However, this lack of scale invariance for both the $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ methods can be remedied quite simply by transforming the data to have identity sample covariance prior to applying either of these constructions. This is a neat result in that it aligns with the recommendations of Wand and Jones (1993), especially when considering a scalar bandwidth matrix. Also, the results in Figure 4.2 that showed $\text{CVBF}_K(\mathcal{S})$ and $\text{CVBF}_K(\mathcal{D})$ performing similarly were artificially optimistic due to the simple covariance matrices. The distributions used for Figure 4.4 are more realistic in practice, and we see the potential benefits for implementing the $\text{CVBF}_K(\mathcal{D})$ approach. However, a simple transformation of

the data allows us to use the simpler and easier to compute $\text{CVBF}_K(\mathcal{S})$ approach, which meets our overall goal for the goodness-of-fit test. An added bonus is the fact that the simpler approach also tends to perform better compared to the slower and slightly more complicated $\text{CVBF}_K(\mathcal{D})$ method.

4.5 Choosing Training Set Size m and Number of Splits N

In order to use any of the multivariate kernel CVBF methods in practice, we must choose values for m and N. Calibration gives us a way of selecting a suitable training set size such that the kernel CVBF performance is appropriate for data from the null model while optimizing the performance for non-null data. The steps for choosing m in the univariate kernel CVBF approach of Hart and Choi (2016) are detailed in Chapter 2, but in the multivariate case we prefer to use a slight modification of these steps.

Methods for choosing the number of random splits to take is not so clear. Of course, in an ideal case, we would take all possible random splits of the data. However, unless n is very small, this number of splits is impossible to consider in practice. Therefore, we must choose N large enough to get a more precise CVBF value, but at the same time, small enough to make computations more efficient. In the univariate case, Hart and Choi (2016) use $N = 100$ in their simulations, but we only use $N = 30$ random splits in the multivariate normality simulations in Sections 4.3 and 4.4. We claimed in Section 4.1 that between 30 and 50 splits will be more than sufficient, so how did we come to this conclusion?

4.5.1 Calibration Steps to Choose m

In order to choose the training set size for a given set of multivariate data, we propose using the following calibration scheme. The overall idea is to compare the observed scaled $\text{CVWE}_K(\mathcal{S})$ curve to the scaled $\text{CVWE}_K(\mathcal{S})$ curve we would expect to see if the null model was the true density function. Looking at these two curves over many training

set sizes, we can choose m such that the CVWE value for the observed data is (nearly) maximized, yet under the null it is as small as possible.

1. Transform the data such that $\hat{\mathbf{\Sigma}} = \mathbf{I_d}$.

2. For $30 \leq N \leq 50$, compute $\text{CVWE}_K(\mathcal{S})$ for training sample sizes $m = \{\lfloor .05n \rfloor, \lfloor .05n \rfloor + 1, \ldots, \lceil .5n \rceil - 1, \lceil .5n \rceil\}$.

3. For at least 25 independent random samples from the null model, compute the scaled $\text{CVWE}_K(\mathcal{S})$ values for the same N and m as in (2). Note: for computational efficiency, choosing fewer training set sizes and $N < 30$ will suffice.

4. On the same set of axes, plot the pairs (m, $\text{CVWE}_{m,N}$) from (2) along with the median, first quartile, and third quartile of the $\text{CVWE}_{m,N}$ values from (3) across the $25+$ samples.

5. Choose $\hat{m}$ as small as possible such that the observed CVWE curve is nearly maximized and the distribution of the null CVWE values is well below 0.

6. If $\text{CVWE}_{\hat{m},N} > \log(3)$ for the observed data, based on Kass and Raftery (1995), conclude that there is at least positive evidence *against* the null model.

7. If $\text{CVWE}_{\hat{m},N} < -\log(3)$ for the observed data, conclude that there is at least positive evidence *in favor* of the null model.

8. If $-\log(3) \leq \text{CVWE}_{\hat{m},N} \leq \log(3)$, there is not enough evidence to favor either model.

The choice of $\hat{m}$ falls under the "art" side of statistics. While we would like to take $\hat{m}$ to be the training set size that maximized the observed CVWE curve, this is not always the best approach. In fact, we will see this in the data analysis of Section 4.9. There

is the interplay between taking $\hat{m}$ to ensure the scaled $\text{CVBF}_K(\mathcal{S})$ method behaves appropriately under the null while optimizing its performance on the observed data. However, the conclusions are the same regardless of the actual CVWE values provided that we remain in the same range using the scale in Table 1.2. For instance any training set sizes that produce CVWE values greater than $\log(150)$ all indicate very strong evidence against the null model. Therefore, we can take $\hat{m}$ to be the smallest value of m such that $\text{CVWE}_{m,N} > \log(150)$.

If we were to reach step (8) in a real data analysis, it would sometimes still be fairly evident based on the plot in step (4) which of the two models is more plausible. If the two CVWE curves are very close and both negative then the null model is certainly plausible. Whereas, if the observed curve is barely positive near its maximum and the null curve is negative, then the alternative model is plausible. We simply cannot make a definitive conclusion in these two cases. For an example of how to implement these calibration steps in practice, see the real data analysis in Section 4.9.

4.5.2 Number of Splits N

In order to determine the plausible number of splits we should consider, we examine how the spread of $\text{CVWE}_K(\mathcal{S})$ values changes as we increase the number of splits. For each of the four distributions (normal, skew-normal, t_3, and Laplace) used in Subsection 4.3.3, we draw 200 random samples of size $n = 500$ from their two- and three-dimensional counterparts (i.e. the skew normal model has parameters $\xi = \mathbf{0}$, $\mathbf{\Omega} = \mathbf{I}_d$, and $\alpha = \mathbf{10}$ regardless of the dimension). The training set size is fixed at $m = 150$ and the numbers of splits evaluated are $N = 1, 2, 5, 10, 15, 20, 25, 30, 40, 60, 80,$ and 100. For each N, we compute the interquartile range of the 200 $\text{CVWE}_{150,N}$ values. Figure 4.5 shows the effect of increasing N on the interquartile range for each of the four distributions in two-dimensions (left panel) and three-dimensions (right panel).

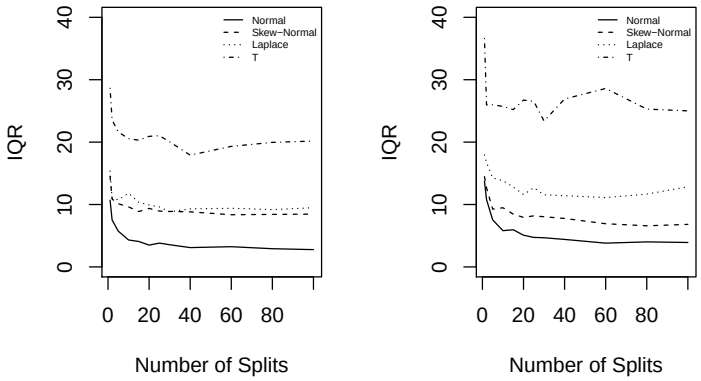

Figure 4.5: Effect of the number of splits on the interquartile range of 200 $\text{CVWE}_K(\mathcal{S})$ values for bivariate data (*left panel*) trivariate data (*right panel*) from a standard normal distribution (solid), t_3 distribution (dashed), skew-normal distribution (dotted), and Laplace distribution (dotdashed).

Regardless of the data dimension, the interquartile range of the $\text{CVWE}_K(\mathcal{S})$ values tends to level off (if not increase slightly) after $N = 20$. It should be noted that similar results hold when using the $\text{CVBF}_K(\mathcal{D})$ method as well as the $\text{CVBF}_K(\mathcal{S})$ approach after re-scaling the observed data vectors. Therefore, taking $30 \leq N \leq 50$ seems reasonable in practice. Using any more splits is not worth the extra computation time.

4.6 Bayes Factor Consistency and Computation in Large Samples

Thus far, we have only considered multivariate data sets of small or moderate size. As mentioned in Chapter 1, one important property of any model comparison technique based on Bayes factors is Bayes factor consistency (Definition 1). In this section, we examine consistency of the $\text{CVBF}_K(\mathcal{S})$ method as the sample size increases under both the null and alternative hypotheses. We will provide sufficient conditions for consistency in Theorems 4.1 and 4.2 with proofs, as well as empirical results that indicate consistency does hold at the more optimal exponential rate for the $\text{CVBF}_K(\mathcal{S})$ method when both the null and

alternative hypotheses are true.

As $n \to \infty$ however, computation of the kernel density estimate becomes increasingly burdensome. Thus, even with a scalar bandwidth matrix, application of the $\text{CVBF}_K(\mathcal{S})$ method to a data set with tens or hundreds of thousands of observations becomes tedious. However, we can take advantage of the consistency results and employ a *Divide and Conquer* strategy to minimize this computation burden.

4.6.1 Mathematical Justification for Consistency

In this subsection, we will show that Definition 1 holds for the $\text{CVBF}_K(\mathcal{S})$ method and that the rate of convergence is exponential under both hypotheses. Note that both Theorems 4.1 and 4.2 only consider a single random split of the data. Therefore, since the performance of the $\text{CVBF}_K(\mathcal{S})$ will only improve for larger N, these consistency results hold for any number of splits.

Under both the null and alternative hypotheses, we use the following notation and make the following assumptions.

Notation:

- Let $X_1, \ldots, X_n \in \mathbb{R}^d \overset{iid}{\sim} f_0$ and $\boldsymbol{X}^T = (X_1, \ldots, X_m)$ and $\boldsymbol{X}^V = (X_{m+1}, \ldots, X_n)$ denote the training and validation sets, respectively.
- Let $\hat{f}_d(\cdot|\boldsymbol{X}^T, h)$ denote the d-dimensional kernel density estimate with scalar bandwidth parameter $h > 0$ and Gaussian kernel function.
- Let $f_d(\cdot|\theta)$ denote the parametric model with parameter $\theta \in \Theta$.

Assumptions:

1. The true parameter value is θ_0 under H_0 such that $f_0 \equiv f_d(\cdot|\theta_0)$. The integral $\int \log f(\mathbf{x}|\theta) f_0(\mathbf{x}) d\mathbf{x}$ exists for all $\theta \in \Theta$, and under the alternative is maximized for some $\theta_0 \in \Theta$.

2. The null marginal likelihood is asymptotic to the Laplace approximation in (4.10) given by

$$(2\pi)^{p/2}(n-m)^{-p/2}|I(\hat{\theta})|^{-1/2}\pi(\hat{\theta})L(\mathbf{X}^V|\hat{\theta}),$$

where $\hat{\theta}$ is the MLE from $\mathbf{X}^V$ and $I(\hat{\theta})$ is the observed information matrix.

3. The MLE $\hat{\theta}$ converges to θ_0 in probability as $n \to \infty$, $I(\cdot)$ and $\pi(\cdot)$ are continuous at θ_0, and $\pi(\theta) > 0$ in a neighborhood of θ_0.

In order to show consistency at an exponential rate under the null hypothesis, we require many additional assumptions. As we will see, the crux of the proof is the Kullback-Leibler discrepancy between $\hat{f}_d(\cdot|h, \boldsymbol{X}^T)$ and $f_d(\cdot|\theta_0)$, which will converge to 0 under the null hypothesis, even though it is strictly positive. We need a number of assumptions to ensure that other terms in the log Bayes factor to tend to 0 at a faster rate than the Kullback-Leibler discrepancy. The assumptions we provide are sufficient conditions for which consistency holds at an exponential rate.

That being said, we know that when the null model is true, we fully expect the parametric model to outperform the kernel model even as $n \to \infty$. In fact, we have seen and will see that empirically, consistency is easier to show under the null than under the alternative. Ironically, the opposite is true mathematically. We now state a consistency result under the null as a theorem, and provide a proof of the result.

Theorem 4.1. *In addition to Assumptions 1-3, also assume the following:*

4. $\frac{1}{n-m}\sum_{j=m+1}^{n}\log f_d(X_j|\theta_0) - \frac{1}{n-m}\sum_{j=m+1}^{n}\log f_d(X_j|\hat{\theta}) = O_p\Big(\frac{1}{\sqrt{n-m}}\Big)$.

5. *Define*

$$D_{KL}(f_d(\cdot|\theta_0), \hat{f}_d(\cdot|h_0, \boldsymbol{X}^T)) = \int \log \frac{f_d(\mathbf{x}|\theta_0)}{\hat{f}_d(\mathbf{x}|h_0, \boldsymbol{X}^T)} f_d(\mathbf{x}|\theta_0)d\mathbf{x}.$$

There exists $h_0 > 0$ *that maximizes* $\int \log \hat{f}_d(\mathbf{x}|h_0, \boldsymbol{X}^T) f_d(\mathbf{x}|\theta_0) d\mathbf{x}$*, and*

$$\int [\log f_d(\mathbf{x}|\theta_0)]^2 f_0(\mathbf{x}) d\mathbf{x} < \infty.$$

Also, $m^\gamma D_{KL}$ *converges in probability to a constant* $C > 0$ *for some* $0 < \gamma < 1$*.*

6. *There exists* $\hat{h} > 0$ *that maximizes the kernel likelihood* $L(\boldsymbol{X}^V|\boldsymbol{X}^T, h)$*.*

7. *For all* $\mathbf{x}, m$*,* $E\big(\log \hat{f}_d(\mathbf{x}|h_0, \boldsymbol{X}^T)\big)^2 \leq g(\mathbf{x})$ *with* $\int g(\mathbf{x}) f_d(\mathbf{x}|\theta_0) d\mathbf{x} < \infty$*.*

8. *Let* $\log \hat{f}_d(\mathbf{x}|\hat{h}, \boldsymbol{X}^T)$ *admit the Taylor series expansion*

$$\log \hat{f}_d(\mathbf{x}|\hat{h}, \boldsymbol{X}^T) = \log \hat{f}_d(\mathbf{x}|h_0, \boldsymbol{X}^T) + (\hat{h} - h_0) \frac{\frac{\partial}{\partial h} \hat{f}_d(\mathbf{x}|h, \boldsymbol{X}^T)|_{\tilde{h}}}{\hat{f}_d(\mathbf{x}|\tilde{h}, \boldsymbol{X}^T)}$$

   where $\tilde{h}$ *is between* h_0 *and* $\hat{h}$*.*

9. *Define* $\hat{g}_l(\mathbf{x}|h, \boldsymbol{X}^T),\ l = 1, \ldots, d$ *to be a multivariate kernel density estimate given by*

$$\hat{g}_l(\mathbf{x}|h, \boldsymbol{X}^T) = \frac{1}{mh^d} \sum_{i=1}^{m} L_l(h^{-1}(\mathbf{x} - X_i)),$$

   where $L_l(h^{-1}(\mathbf{x} - X_i)) = \frac{(x_l - X_{il})^2}{h^2} K_d(h^{-1}(\mathbf{x} - X_i))$ *for Gaussian kernel function* $K_d(\cdot)$*. Then,*

$$\frac{1}{n-m} \sum_{j=m+1}^{n} \frac{\hat{g}_l(X_j|\tilde{h}, \boldsymbol{X}^T)}{\hat{f}_d(X_j|\tilde{h}, \boldsymbol{X}^T)} - 1 = O_p(1), \quad l = 1, \ldots, d.$$

10. *For some* $0 < a < 1$*,* $(\hat{h} - h_0)/\tilde{h} = O_p(n^{-a})$*.*

11. *The training set size* m *is such that* $m = n^b$ *with* $b < \min\left(\frac{a}{\gamma}, \frac{1}{2\gamma}\right)$*.*

Under Assumptions 1-11

$$\log \boldsymbol{BF}_{m,1} \leq -nm^{-\gamma}C + o_p(nm^{-\gamma}).$$

Proof of Theorem 4.1 Using the Laplace approximation for the null marginal, we can bound the $\log \text{BF}_{m,1}$ value for a scalar bandwidth parameter by

$$\begin{aligned}\log \text{BF}_{m,1} \leq & \sum_{j=m+1}^{n} \log \hat{f}_d(X_j|\hat{h}, \boldsymbol{X}^T) - \sum_{j=m+1}^{n} \log f_d(X_j|\theta_0) \\ & + \sum_{j=m+1}^{n} \log f_d(X_j|\theta_0) - \sum_{j=m+1}^{n} \log f_d(X_j|\hat{\theta}) + O_p(\log n).\end{aligned}$$

From Assumptions 4 and 5, the fact that

$$\frac{1}{n-m} \sum_{j=m+1}^{n} \log \hat{f}_d(X_j|\hat{h}, \boldsymbol{X}^T) = \int \log \hat{f}_d(\mathbf{x}|h_0, \boldsymbol{X}^T) f_d(\mathbf{x}|\theta_0) d\mathbf{x} + \delta_1 + \delta_2,$$

where

$$\begin{aligned}\delta_1 &= \frac{1}{n-m} \sum_{j=m+1}^{n} \log \hat{f}_d(X_j|\hat{h}, \boldsymbol{X}^T) - \frac{1}{n-m} \sum_{j=m+1}^{n} \log \hat{f}_d(X_j|h_0, \boldsymbol{X}^T), \\ \delta_2 &= \frac{1}{n-m} \sum_{j=m+1}^{n} \log \hat{f}_d(X_j|h_0, \boldsymbol{X}^T) - \int \log \hat{f}_d(\mathbf{x}|h_0, \boldsymbol{X}^T) f_d(\mathbf{x}|\theta_0) d\mathbf{x},\end{aligned}$$

and noting that

$$\frac{1}{n-m} \sum_{j=m+1}^{n} \log f_d(X_j|\theta_0) = \int \log f_d(\mathbf{x}|\theta_0) f_d(\mathbf{x}|\theta_0) d\mathbf{x} + O_p\Big(\frac{1}{\sqrt{n-m}}\Big),$$

we can write

$$\log \text{BF}_{m,1} \leq (n-m)m^{-\gamma}[-m^{\gamma} D_{KL}(f_d(\cdot|\theta_0), \hat{f}_d(\cdot|h_0, \boldsymbol{X}^T)) + m^{\gamma}\delta_1 + m^{\gamma}\delta_2] + O_p(\sqrt{n-m}).$$

By Assumptions 5 and 11, the result will hold provided that $m^{\gamma}\delta_i \xrightarrow{p} 0$ for $i = 1, 2$. From the Taylor series expansion in Assumption 8, we can write δ_1 as

$$\delta_1 = \frac{1}{n-m}(\hat{h} - h_0) \sum_{j=m+1}^{n} \frac{\frac{\partial}{\partial h}\hat{f}_d(X_j|h, \boldsymbol{X}^T)|_{\tilde{h}}}{\hat{f}_d(X_j|\tilde{h}, \boldsymbol{X}^T)}.$$

Using the kernel estimate from Assumption 9, the derivative of the kernel density estimate is given by

$$\frac{\partial}{\partial h}\hat{f}_d(X_j|h, \boldsymbol{X}^T)|_{\tilde{h}} = \frac{1}{\tilde{h}}\Big[\sum_{l=1}^{d} \hat{g}_l(X_j|\tilde{h}, \boldsymbol{X}^T) - d\hat{f}_d(X_j|\tilde{h}, \boldsymbol{X}^T)\Big].$$

Therefore,

$$\delta_1 = \left(\frac{\hat{h} - h_0}{\tilde{h}}\right) \sum_{l=1}^{d} \left(\frac{1}{n-m} \sum_{j=m+1}^{n} \frac{\hat{g}_l(X_j|\tilde{h}, \mathbf{x}^T)}{\hat{f}_d(X_j|\tilde{h}, \mathbf{x}^T)} - 1\right).$$

Thus, by Assumptions 9 and 10, $\delta_1 = O_p(n^{-a})$ and so $m^{\gamma}\delta_1 \xrightarrow{p} 0$ since $b < a/\gamma$.

Next, consider δ_2, which is completely determined by the training data. Therefore, conditioning on $\boldsymbol{X}^T$ and using the fact that

$\mathrm{E}[\log \hat{f}_d(X_j|h_0, \boldsymbol{X}^T)|\boldsymbol{X}^T] = \int \log \hat{f}_d(\mathbf{x}|h_0, \boldsymbol{X}^T) f_d(\mathbf{x}|\theta_0) d\mathbf{x}$ for $j = m+1, \ldots, n$,

$$\begin{aligned}
\mathrm{E}[\delta_2^2] &= \mathrm{E}[\mathrm{E}(\delta_2^2|\boldsymbol{X}^T)] \\
&= \mathrm{E}\Big[\mathrm{Var}\Big(\frac{1}{n-m} \sum_{j=m+1}^{n} \log \hat{f}_d(\boldsymbol{X}_j|h_0, \boldsymbol{X}^T)\Big|\boldsymbol{X}^T\Big)\Big] \\
&= \frac{1}{n-m}\mathrm{E}\Big[\int (\log \hat{f}_d(\mathbf{x}|h_0, \boldsymbol{X}^T))^2 f_d(\mathbf{x}|\theta_0) d\mathbf{x} - \Big(\int \log \hat{f}_d(\mathbf{x}|h_0, \boldsymbol{X}^T) f_d(\mathbf{x}|\theta_0) d\mathbf{x}\Big)^2\Big].
\end{aligned}$$

Thus, by Assumption 7, $\delta_2 = O_p\Big(\frac{1}{\sqrt{n-m}}\Big)$, which means, $m^{\gamma}\delta_2 \xrightarrow{p} 0$ and hence, we have reached the desired result.

We should point out that many of the assumptions in Theorem 4.1 are made to expedite the proof. However, all are reasonable based on a combination of intuition and known

results. For instance, Assumption 9 is anticipated in light of Hall (1987), Hall and Marron (1987), and van der Laan et al. (2004).

Now we consider the scenario where the kernel model is true. We take a similar approach to Hart and Choi (2016) in that we approximate the alternative marginal likelihood using a quadrature approximation, namely a Riemann sum over a finite support. In practice, we noted that a Laplace approximation is our preferred method for computing the marginal. However, asymptotically, the bandwidth parameter approaches a boundary point, which is problematic for applying maximum likelihood methods in the Laplace approximation. That being said, any quadrature approach works very reliably in the scalar bandwidth case, albeit a little slower computationally depending on how many evaluation points we consider. Similar to the null case, we state the additional assumptions and result in Theorem 4.2, followed by a proof.

Theorem 4.2. *In addition to Assumptions 1-3, also assume the following:*

12. *The alternative marginal likelihood can be approximated by the Riemann sum approximation given by*

$$\frac{h_M - h_1}{M-1} \sum_{k=1}^{M} L(\boldsymbol{X}^V | h_k, \boldsymbol{X}^T) p(h_k)$$

where the set of evaluation points $\{h_1, \ldots, h_M\}$ *are equally spaced such that* $h_M = m^{-\beta}$, $h_1 = m^{-\alpha}$, $h_k = h_1 + (h_M - h_1)(k-1)/(M-1),\ k = 1, \ldots, M$, $0 < \beta < \alpha$, $1/4 < \alpha < 1$, *and* $m = o(n)$ *for some arbitrarily large* n.

13. *The quantity* $\int [\log f_0(\mathbf{x})]^2 f_0(\mathbf{x}) d\mathbf{x}$ *is finite.*

14. *The kernel likelihood evaluated at* h_1, $\frac{1}{n-m} \sum_{j=m+1}^{n} \log \hat{f}_d(X_j | h_1, \boldsymbol{X}^T)$, *is consistent for* $\int \log f_0(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x}$.

15. *The training set size* $m = n^c$ *where* $0 < c < 1$.

16. *The quantity* $\log\left(m^{-1}\sum_{i=1}^{m}\gamma_i^{-(d-1)/2}\right) = o_p(n)$ *where* $\gamma_i = \frac{1}{2}[\mathbf{w} - X_i]^T[\mathbf{w} - X_i]$ *in (4.3) and* $\mathbf{w}$ *is the vector of column medians from* $\boldsymbol{X}^V$.

17. *The true density function is continuous in a neighborhood of* $\mathbf{w}$ *and* $f_0(\mathbf{w}) > 0$.

If $m \to \infty$*, then as* $n \to \infty$ *the approximate Bayes factor for a single random split*

$$\tilde{BF}_{m,1} = \frac{\frac{h_M - h_1}{M-1}\sum_{k=1}^{M} L(\boldsymbol{X}^V|h_k, \boldsymbol{X}^T)p(h_k)}{(2\pi)^{p/2}(n-m)^{-p/2}|I(\hat{\theta})|^{-1/2}\pi(\hat{\theta})L(\mathbf{X}^V|\hat{\theta})}$$

is bounded below by

$$\exp\big(nD_{KL}(f_0(\cdot), f_d(\cdot|\theta_0)) + o_p(n)\big).$$

Proof of Theorem 4.2. Noting that

$$\sum_{k=1}^{M} L(\boldsymbol{X}^V|h_k, \boldsymbol{X}^T)p(h_k) \geq L(\boldsymbol{X}^V|h_1, \boldsymbol{X}^T)p(h_1),$$

with probability tending to 1, the approximate Bayes factor is at least

$$\begin{aligned}\tilde{\text{BF}}_{m,1} \geq\ & B(n-m)^{-p/2}\frac{h_M - h_1}{M}p(h_1) \\ & \times \exp\Big(\sum_{j=m+1}^{n}\log\hat{f}_d(\boldsymbol{X}_j|h_1, \boldsymbol{X}^T) - \sum_{j=m+1}^{n}\log f_d(X_j|\hat{\theta})\Big),\end{aligned}$$

where B is a positive constant. Taking the log Bayes factor, we can write this inequality as,

$$\begin{aligned}\log\tilde{\text{BF}}_{m,1} \geq (n-m)\Bigg(&\frac{1}{n-m}\sum_{j=m+1}^{n}\log\hat{f}_d(\boldsymbol{X}_j|h_1, \boldsymbol{X}^T) - \frac{1}{n-m}\sum_{j=m+1}^{n}\log f_d(X_j|\hat{\theta}) \\ &+ \int\log f_d(\mathbf{x}|\theta_0)f_0(\mathbf{x})d\mathbf{x} - \int\log f_d(\mathbf{x}|\theta_0)f_0(\mathbf{x})d\mathbf{x}\Bigg) + \log p(h_1) + O(\log n).\end{aligned}$$

Using Assumptions 12-14,

$$\tilde{\text{BF}}_{m,1} \geq (n-m)\Big(D_{KL}(f_0(\cdot), f_d(\cdot|\theta_0)) + \frac{1}{n-m}\log p(h_1) + o_p(1) + O\Big(\frac{\log n}{n}\Big)\Big).$$

The result will be shown provided that $\frac{1}{n-m}\log p(h_1) \to 0$. Taking the log of the prior distribution in (4.3) evaluated at h_1,

$$\log p(h_1|\gamma) = \log(A) + \log(\hat{f}_d(\mathbf{w}|\boldsymbol{X}^T, h_1)) - \log\Big(m^{-1}\sum_{i=1}^{m}\gamma_i^{-(d-1)/2}\Big),$$

where A is a positive constant. Since $\hat{f}_d(\mathbf{w}|\boldsymbol{X}^T, h_1)) \xrightarrow{p} f_0(\mathbf{w})$, it follows from Assumptions 16 and 17 that $\log p(h_1)/n = o_p(1)$. Therefore, we have reached the desired result.

There are a few important details to point out in this theorem and proof. First, Assumption 14 is included out of necessity, but it is not unreasonable since as $n \to \infty$ we know that the kernel density estimate evaluated at the smallest bandwidth evaluation point h_1 will converge in probability to the true density function so long as $h_1 \to 0$ with $mh_1 \to \infty$. The remaining assumptions (save for Assumption 12) are rather weak, but are included since a few select densities could prove to be problematic. For instance, one could construct a bimodal density such that $f_0(\mathbf{w}) = 0$. In this case, $\log p(h_1) = -\infty$ and the consistency result would not hold without Assumption 17.

4.6.2 Empirical Consistency Results

While Theorems 4.1 and 4.2 require many assumptions, they do indicate that an exponential rate of consistency can be attained under both hypotheses as the sample size increases towards infinity. In order to verify that consistency of the $\text{CVBF}_K(\mathcal{S})$ method holds for more practical sample sizes ($n \leq 10,000$), consider the following small simulations. For sample sizes $n = 500, 1000, 2000, 5000,$ and 10000, 32 independent random samples are drawn from four-dimensional distributions either from the null model

(standard normal distribution) or the alternative model (Laplace and skew-normal distributions). For each of these random samples, we compute the scaled $\text{CVWE}_K(\mathcal{S})$ value using training set proportions $p = .1, .2, .3, .4,$ and $.5$ and $N = 30$ random splits. A training set proportion p simply means that we randomly split the data such that $m = pn$ observations are in the training set.

Figure 4.6 displays the resulting scaled $\text{CVWE}_K(\mathcal{S})$ values when we assume the null model to be true and sample data from the standard normal distribution. Notice that as the sample size increases, the entire CVWE curve shifts toward $-\infty$. Each curve has the same rough shape, monotonically increasing as the training set size increases, that we have come to expect under the null model. Also, as a function of sample size, for a fixed training set proportion, the decrease in CVWE values is nearly linear. Though we do not include the results here, we have found a similar relationship between sample size and CVWE values for two- and three-dimensional normal data. Therefore, these empirical results lend credence to the Kullback-Leibler discrepancy being the dominant term in the log Bayes factor leading us to conclude that the $\text{CVBF}_K(\mathcal{S})$ method is consistent under the null hypothesis at an exponential rate.

Figure 4.7 contains the simulation results when the alternative model is true. The top and bottom panels correspond to the scaled $\text{CVWE}_K(\mathcal{S})$ values when the data are sampled from the skew-normal $(SN(\mathbf{0}, \mathbf{I}_4, \mathbf{10}))$ and Laplace distributions (each coordinate $L(0, 1)$), respectively. Unlike Figure 4.6 where the conclusions regarding normality were the same at every sample size and training set size, under the alternative models things are not so clean. For the skew-normal data, because we need a larger number of observations in the training set for the kernel model to be accurate, the consistency results are less obvious. For instance, when $m = .4n$, there is a slow but steady increase in CVWE values for the smaller sample sizes $n \leq 2000$. This increase becomes more apparent when $n \geq 5000$. Now consider the smallest training set size of $m = .1n$. As the sample size

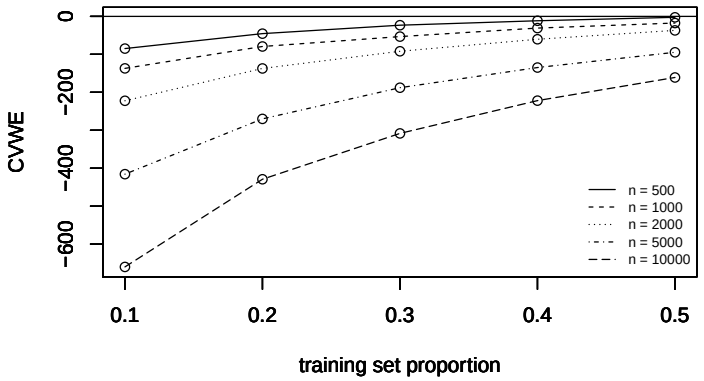

Figure 4.6: Bayes factor consistency of the scaled $\text{CVBF}_K(\mathcal{S})$ ($N = 30$, $p = .1, .2, .3, .4, .5$) method when testing four-dimensional normality for standard normal data. In decreasing order, the curves correspond to the following sample sizes: $n = 500, 1000, 2000, 5000$ and 10000.

increases to $n = 2000$, we actually see a slight *decrease* in CVWE values. Once the sample size increases beyond $n = 2000$, so that the training set contains enough observations, however, the CVWE values increase as expected. As for the Laplace data, consistency is much clearer to see. The CVWE curves slowly shift upwards away from 0 for sample sizes $n = 500, 1000,$ and 2000. For sample sizes $n > 2000$, the curves shoot off toward ∞.

The most important takeaway from Figure 4.7 is the lack of agreement in the respective conclusions against/for normality across all sample sizes for both distributions. When $n = 500$, we would conclude that the Laplace data were normally distributed when $m = 100, 200$ (similarly for $n = 1000$, $m = 200$). As for the skew-normal model and any combination of n and p such that $m \leq 500$, the scaled $\text{CVWE}_K(\mathcal{S})$ values will favor normality. Since we know the true data generating distributions in this simulation, we know that we are incorrectly favoring normality in these instances. However, in practice,

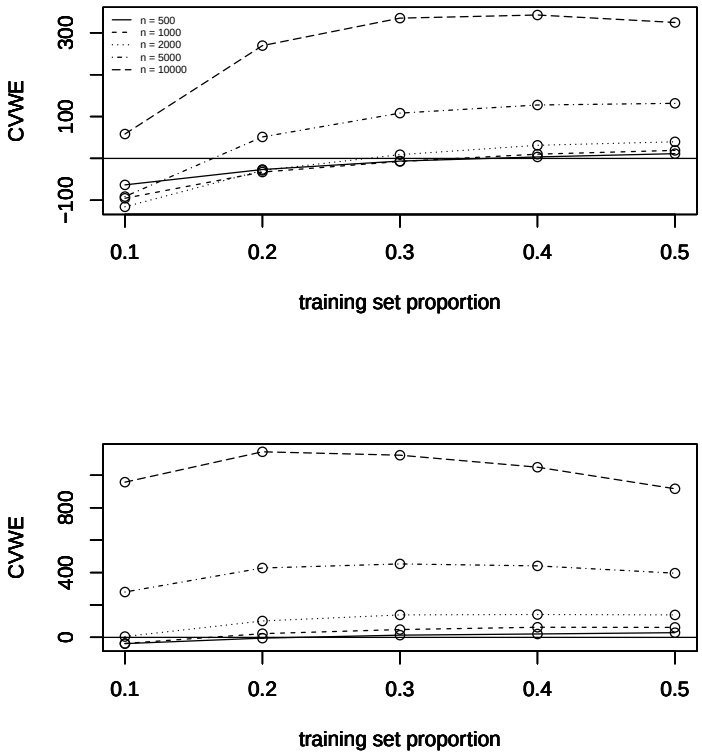

Figure 4.7: Bayes factor consistency of the scaled $\text{CVBF}_K(\mathcal{S})$ ($N = 30$, $p = .1, .2, .3, .4, .5$) method when testing four-dimensional normality for skew-normal data (*top panel*) and Laplace data (*bottom panel*). Each curve corresponds to one of the following sample sizes: $n = 500, 1000, 2000, 5000$ and 10000.

we do not know the true density function. Therefore, when the sample size is small for multivariate data, it is imperative to implement the calibration scheme of Subsection 4.5.1 to compare the observed CVWE curve to the null curve. That being said, consistency does hold at an exponential rate so once we have a large enough sample, the probability of correctly favoring the kernel model tends to 1 for any alternative model.

4.6.3 Divide and Conquer Kernel CVBF

The kernel density estimator in (3.1) with Gaussian kernel function is inefficient to compute for large sample sizes as described in Raykar et al. (2010). In order to compute the likelihood function $L(\mathbf{X}^V|\mathbf{H}, \mathbf{X}^T)$, $m(n-m)$ evaluations of the kernel function are required for a single random split. This scales quadratically with increasing sample size, which becomes incredibly costly. For an example of how computation time scales with increasing sample size, consider computing the scaled CVWE($\mathcal{S}$) value for a single four-dimensional t_3 random sample with 30 random splits and training sample proportion $p = .30$. On a MacBook Pro (2.8 GHz Intel Core i5, 16 GB RAM) the respective run times for sample sizes $n = 500, 1000, 2500, 5000,$ and 10000 are 3.62, 7.32, 25.80, 125.13, and 583.39 seconds.

In an effort to decrease the computational burden, we could implement one of the many approximation methods to greatly reduce the number of kernel function evaluations required to compute the kernel likelihood function. Most of these approximations utilize some form of *binning* (Silverman (1986), Härdle and Scott (1992), Wand (1994), and Tang and Karunamuni (2016)). When applied to multivariate kernel density estimation, let $g_{l1} < \cdots < g_{lM}$ be an equally spaced grid of $M << n$ points ($M_l = M$ for simplicity) in the l-th coordinate direction for $l = 1, \ldots, d$ such that all observed data values are contained within the grid. The raw observations are then replaced with grid counts using a binning rule like simple, linear, centered, or rounding. These grid counts represent the amount of data within a neighborhood of a given grid point. In their simplest form, the binned approximation to the kernel likelihood function now requires $O(M^{2d}(n-m))$ kernel evaluations (Wand, 1994), which is only more efficient when M is very small compared to n. Utilization of the fast Fourier transform can reduce the number of kernel evaluations to a much more efficient $O((M\log(M))^d(n-m))$. Keep in mind that increasing M will

lead to more accurate approximations, while at the same time, increasing the computation time required (especially in higher dimensions). Even though these binning approximations may be more efficient than the brute force computation of the kernel likelihood, we would still see an increase in run time as the sample size gets very large. Thus, how can we expedite the computation of kernel CVBF, regardless of computation method, without compromising the overall conclusions?

A common approach for tackling massive data sets is known as *Divide and Conquer* and has been discussed by many authors (see Li et al. (2013) and Chang et al. (2016) for a couple examples). The idea is very simple. Suppose we want to compute a statistic $\hat{\theta}_n$ from the entire sample $Z_1, Z_2, \ldots, Z_n$ where n is extremely large. In an effort to save run time and computer memory, randomly partition the sample into w smaller data sets of equal size k. Now, compute the statistic of interest on each of the partitions to obtain $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_w$. Finally, recombine the w statistics, each based on k observations, appropriately to find $\tilde{\theta}$, such that $\tilde{\theta}$ is comparable to the overall statistic $\hat{\theta}_n$. Bhattacharya and Hart (2016) apply this idea to determining the optimal bandwidth parameter in kernel density estimation using partitioned cross-validation. They partition the data into disjoint subgroups, find the optimal smoothing parameter for each subgroup using standard cross-validation techniques, and then re-scale and average the smoothing parameters across subgroups to get an overall smoothing parameter. Another common application is in estimating slope parameters in (generalized) linear models for subgroups by simply averaging the estimates to determine the model for the entire dataset (Chang et al., 2016). For more applications of *Divide and Conquer* strategies, see the references within Bhattacharya and Hart (2016).

Does a *Divide and Conquer* approach make sense in the kernel CVBF methodology? First, the overall run time for computing the w CVWE values on all partitions must be faster than the run time on the entire data set. Next, the appropriate number of partitions must be used such that the conclusions from each partition are not contradictory to

the overall conclusions. Finally, an appropriate method must be derived such that the recombined estimate from the partitions is comparable to the overall CVWE value. In the discussion to follow, we only consider the scaled $\text{CVBF}_K(\mathcal{S})$ method.

Since the computation time increases quadratically with increasing sample size, the time required to compute the CVWE value over all partitions will certainly be quicker. As a simple example, suppose we had a four-dimensional data set of $n = 10,000$ observations from a t_3 distribution. Using the run times given above for a 30% / 70% split into training and validation set and $N = 30$ random splits, if we partitioned the largest data set into $w = 20, 10, 5,$ and 2 partitions, the overall run times across the partitions would be 72.34, 73.20, 129.00, and 250.26 seconds for subgroup sizes, $k = 500, 1000, 2000,$ and 5000, respectively. All of these times are less than the 583.39 seconds required for one split of the entire dataset. Therefore, in general, the overall run time decreases as the number of partitions increases, however there is a point where the decrease reaches a minimum and then either levels off or begins to increase slightly. This means that there is a computational limit such that the trade-off between the number of partitions and the sample size in each partition produces (roughly) the same overall run time. However, this limit is still a fraction of the computation time required to compute the kernel CVWE value on the entire data set.

In the results of Figure 4.7, we noted that for certain combinations of training set size m and sample size n, the $\text{CVBF}_K(\mathcal{S})$ method incorrectly favored the null model for four-dimensional Laplace and skew-normal data. This poses a problem for applying a *Divide and Conquer* algorithm since the conclusions based on each partition may contradict the overall results if we choose w to be too large. Therefore, choosing w is incredibly important. We want to use as few partitions as possible to ensure that we reach correct conclusions in each partition, but at the same time the number of partitions should be big enough to substantially speed up computations. A practical solution to optimize this trade-

off is to select a variety of w values and then for each w, compute the CVWE$_K(\mathcal{S})$ value (and run time) on a single random sample from the original data set of size $k_w = n/w$. Keep in mind that the training set proportion plays a pivotal role in the choice of w as well. As w increases, k_w decreases, so for a fixed training set proportion, the training set size within a subgroup will decrease. Therefore, for increasing w, we must let p increase toward $.5$ in order for the training set to contain enough observations to produce an accurate kernel density estimate. This idea is backed up in the theory which says the training set proportion tends to 0 as sample size increases. Once we have our test CVWE values for different subgroup sizes, we should be able to get a sense which hypothesis is favored by examining the trend toward $\pm\infty$ as $w \to 1$. We then choose one of the w values for which the CVWE$_K(\mathcal{S})$ conclusions are in agreement, taking into account the respective computation times, as the final number of partitions.

Often times in practice, the above scheme for choosing w is unnecessarily extensive. We can begin by taking random subsamples from the data which are both "large enough" and computationally manageable. Then compute the CVWE$_K(\mathcal{S})$ value using a suitable training set size. If the evidence from that subsample is overwhelmingly in favor (or against) the null model, then we can likely stop with the exploration and report our conclusions. For more satisfactory conclusions, take a new random subsample of the same size and repeat the process. The only time we want to carefully choose w and p is when the amount of evidence is close to 0. However, under consistency, if the subsample size is large enough, for any appropriate training set size, the CVWE$_K(\mathcal{S})$ value will be sufficiently far from 0.

The toughest part of a *Divide and Conquer* algorithm is determining the appropriate method of recombination. In the kernel CVBF method, this is more complicated than a simple average and/or rescaling. In Figures 4.6 and 4.7, not only does the CVBF value shift away from 0 (direction depending on which model is true), the CVWE values are not

exactly a linear function of sample size. Therefore, a scaling value will be some function of w and averaging all of the CVWE$_w$ values will result in an underestimate of the overall CVWE value. However, since the CVBF$_K(\mathcal{S})$ method is consistent at an exponential rate under both hypotheses, if $|\text{CVWE}_w| > q$ for critical value q, then $|\text{CVWE}| > q$ for the entire data set. This means that if we choose w appropriately, our conclusions in terms of the amount of evidence in favor (or against) the null model will agree.

In order to see how a *Divide and Conquer* strategy would apply in the kernel CVBF context, consider the four-dimensional distributions used in the simulation for testing normality in Subsection 4.3.3. For each of the four distributions, we randomly sample a single data set of $n = 10,000$ observations. We then apply a *Divide and Conquer* strategy using $w = 2, 5, 10, 20, 40,$ and 100 subgroups. Note that we would never consider using 100 subgroups with 100 observations each in practice. This row is included simply to show that the computation time will begin to increase for w too large. Within each subgroup, we compute the scaled CVWE$_K(\mathcal{S})$ value with $N = 30$ and $m = .3k$. We mentioned that the training set size should change with w, however for simplicity we keep it fixed at a proportion that is reasonable in all cases considered here. The recombination method is simply the median CVWE value across all w subgroups. Table 4.1 provides the CVWE$_K(\mathcal{S})$ values for each *Divide and Conquer* scheme based on the six subgroup sizes for each data set, the CVWE$_K(\mathcal{S})$ value on each entire data set, and the respective computation times to produce each CVWE$_K(\mathcal{S})$ in the table.

If we were to compute the scaled CVWE$_K(\mathcal{S})$ value from each of the four distributions we certainly find overwhelming evidence in favor of the correct model. However, for 30 random splits and just a single training set size of $m = 3,000$, the computations would take around 8 to 10 minutes. Keep in mind that in practice we would want to repeat these computations for a large number of training set sizes on both the observed data and data sampled from the null model. Therefore a real data analysis would take hours to complete.

$(w,\, k)$	Normal	Skew-Normal	t_3	Laplace
$1\ (10,000)$	$-298.09\ (476)$	$323.48\ (441)$	$1994.04\ (600)$	$1080.04\ (562)$
$2\ (5,000)$	$-182.37\ (237)$	$97.65\ (260)$	$695.74\ (287)$	$433.17\ (310)$
$5\ (2,000)$	$-86.77\ (129)$	$7.76\ (130)$	$180.98\ (128)$	$118.80\ (145)$
$10\ (1,000)$	$-50.90\ (79)$	$-6.47\ (91)$	$56.56\ (103)$	$42.67\ (105)$
$20\ (500)$	$-26.06\ (65)$	$-5.27\ (69)$	$10.24\ (68)$	$16.71\ (66)$
$40\ (250)$	$-10.08\ (62)$	$-2.98\ (69)$	$6.20\ (74)$	$5.71\ (61)$
$100\ (100)$	$0.03\ (84)$	$2.07\ (88)$	$4.74\ (91)$	$4.41\ (72)$

Table 4.1: Application of a *Divide and Conquer* scheme to testing four-dimensional normality of a single data set of $n = 10,000$ observations from a normal, skew-normal, t_3, and Laplace distributions. Each data set is partitioned into $w = 1, 2, 5, 10, 20, 40$, and 100 subgroups and the scaled $\text{CVBF}_K(\mathcal{S})$ method is applied to each partition with $N = 30$ and $m = .3k$. The median $\text{CVWE}_K(\mathcal{S})$ value across all w partitions is reported along with the respective computation times.

To speed up the analysis, we might consider a *Divide and Conquer* approach. Remember that we want to choose w so that we have enough observations k in each subgroup so that our conclusions agree with the entire data set. However, we want w to be small enough so that we minimize the total computation time. First consider the column for the normal data in Table 4.1. We would find strong evidence in favor of the normal model with as many as $w = 40$ subgroups with $k = 250$ observations. The overall computation time to compute the $\text{CVWE}_K(\mathcal{S})$ value of -10.08 is a mere 62 seconds compared to 476 seconds for the entire data set (13% the full run time). Of course, with $k = 250$, the training set within each subgroup is only 75 observations. Therefore, we may consider the extra 17 seconds of computation time to use $w = 10$, which results in a $\text{CVWE}_K(\mathcal{S})$ value of

-50.90, but more importantly each training set contains 300 observations. This is why we recommend determining the computing time and $\text{CVWE}_K(\mathcal{S})$ value for a single subgroup of a few partitions of size w.

For the t_3 and Laplace distributions, we conclude against the normal model in all schemes. However, the overall run times begin to increase once w gets too large. For the t_3 distribution this is at $w = 40$ and for the Laplace distribution when $w = 100$. So with such strong results at $w = 20$ ($\text{CVWE}_K(\mathcal{S})$ values of 10.24 and 16.71 for t_3 and Laplace, respectively) and respective computation times just over one minute, we can strongly favor the kernel model in both cases in about 10% the run time.

Finally, the skew-normal model illustrates the need for choosing w large enough such that the conclusions reached on the subgroups agree with those from the whole data set. When $w \leq 5$, we find strong evidence against normality and when $10 \leq w \leq 40$ we find positive to strong evidence in favor of normality. Therefore, we would likely take $w = 5$, since we have strong evidence against normality ($\text{CVWE}_K(\mathcal{S})$ value of 7.76), but a computation time of 2 minutes (29% of the overall time). We could double our computing time for far stronger results ($\text{CVWE}_K(\mathcal{S})$ value of 97.65), but for the purpose of testing goodness-of-fit, this is unnecessary.

The moral of this simulation, and this section for that matter, is that a *Divide and Conquer* approach could be extremely useful when testing goodness-of-fit with the scaled $\text{CVBF}_K(\mathcal{S})$ method on data with moderate dimension and extremely large sample size. Of course, there is some fine-tuning required in the choice of w as well as training set proportion p before conducting a full analysis. However, once w and p are chosen, simply carry out the calibration steps of Subsection 4.5.1 using the resulting $\text{CVWE}_K(\mathcal{S})$ values from the *Divide and Conquer* scheme. Because the scaled $\text{CVBF}_K(\mathcal{S})$ method is consistent, the resulting conclusions *must* be stronger for the overall data than those from the *Divide and Conquer* approach.

4.7 Comparison to Frequentist Goodness-of-Fit Tests

In Chapter 1, we looked at a variety of Bayesian and frequentist multivariate goodness-of-fit tests and compared the general hypothesis testing framework from both perspectives. We argued that taking a Bayesian approach to hypothesis testing had many advantages, especially in the goodness-of-fit problem. In this section, we look at how the kernel CVBF method compares in performance to the more common frequentist multivariate goodness-of-fit tests in a simple simulation for testing trivariate normality.

We only consider frequentist methods in this simulation because of all the Bayesian tests mentioned in Subsection 1.1.3, the only one capable of testing multivariate normality is that of Tokdar and Martin (2013). However, remember that the motivation for the kernel CVBF method is to have a simple and intuitive Bayesian approach to testing goodness-of-fit for any parametric null model. The method of Tokdar and Martin (2013), while effective for testing multivariate normality, is neither simple nor intuitive. In reality, statisticians verify multivariate normality assumptions using one of the following frequentist tests (Korkmaz et al., 2016): Mardia's test for skewness, Mardia's test for kurtosis, Royston's test, and the Henze-Zirkler test. Therefore, we will compare the $\text{CVBF}_K(\mathcal{S})$, $\text{CVBF}_K(\mathcal{D})$, and the scaled $\text{CVBF}_K(\mathcal{S})$ methods to the four aforementioned frequentist tests in the following simulation.

Consider 1,000 trivariate random samples of size $n = 1,000$ from the normal, Laplace, and skew-normal families, where each sample uses randomly generated parameters. For each data set, the four frequentist tests as well as three kernel CVBF methods at training set sizes $m = 100, 200, 300, 400,$ and 500 are carried out. Using the normal data, we first want to compare the Type I error rates for all the tests considered. Next, we look at a power study using the skew-normal and Laplace data. Not only will these comparisons compare our kernel CVBF methods to common frequentist tests, but we will also be able

to finally recommend a single kernel CVBF method.

4.7.1 Type I Error Rates

For the frequentist goodness-of-fit tests of normality, a Type I error is made when $P < .05$ when the data are truly normally distributed. We will define a Type I error to be $\text{CVWE}_{m,N=32} > \log(3)$, since this represents positive evidence against the normal model according to Kass and Raftery (1995). Table 4.2 provides the number of Type I errors for each of the four frequentist tests.

	Mardia (Skew)	Mardia (Kurtosis)	Royston	Henze-Zirkler
Type I Errors (Rate)	59 (.059)	44 (.044)	80 (.080)	49 (.049)

Table 4.2: Number of Type I errors in 1,000 randomly generated trivariate normal distributions with $n = 1,000$ using common frequentist goodness-of-fit tests for normality.

We expect the Type I error rates for the frequentist tests to be fairly close to $\alpha = .05$. This is true for both of Mardia's tests and the Henze-Zirkler test with Type I error rates ranging from .044 to .059. However, Royston's test produced a surprisingly large Type I error rate of .08. The corresponding Type I error rates for the three kernel CVBF methods are not included in a table since for all but 2 of the 15 tests, zero Type I errors are made. Only two and five errors are made when $m = 500$ for the $\text{CVBF}_K(\mathcal{S})$ approach for the original data and the scale transformed data, respectively. Based on the simulations in this chapter and the steps of calibration, a 50/50 split of the data is less than ideal and would rarely (if ever) be used in practice. Under the null model, the CVBF curve increases monotonically to 0 and thus we expect poorer performance for normal data when $m = 500$. Also, as $n \to \infty$ the consistency results in Section 4.6 and in Hart and Choi (2016) require

$m/n \to 0$, which means the training set size will be a smaller proportion of the data as n increases. Even if we relax the definition of a Type I error under the kernel CVBF methods to be $\text{CVWE}_{m,N=32} > -\log(3)$, then we still fail to make a Type I error provided $m \leq 400$. This means that in all 1,000 samples, we have positive evidence in favor of normality.

An interesting question to ask is what the level of each frequentist test should (roughly) be in order for the Type I error rates to agree with the kernel CVBF methods. Since we did not make any Type I errors for suitable choices of m, we cannot find the corresponding level using this simulation. Conservatively though, we could set $\alpha = .001$, which implies the frequentist tests make a Type I error in 1 out of every 1,000 tests. This significance level is far less than the usual .05 level. In frequentist testing, the probability of a Type I error remains fixed at the chosen level α as $n \to \infty$. However, we know from the consistency of the $\text{CVBF}_K(\mathcal{S})$ method that as $n \to \infty$, $P(\text{Type I Error}) \to 0$ and thus the level of the frequentist test should in fact tend to 0 as well.

4.7.2 Power Study

Now that we have compared the performance of kernel CVBF methods to frequentist tests of multivariate normality when the null hypothesis is true, it makes sense to see what happens when the alternative hypothesis is true. The empirical power of a test γ is defined as the probability of correctly concluding in favor of the alternative hypothesis. For the kernel CVBF methods, we find positive evidence in favor of the kernel model when $\text{CVWE}_{m,N=32} > \log(3)$. Instead of using $\alpha = .05$ for the frequentist tests, to make the comparisons fair we use $\alpha = .001$ since the corresponding Type I error rates are now roughly equivalent across all tests. Tables 4.3 and 4.4 contain the respective powers of each test for skew-normal and Laplace data. Notice that we do not include the kernel CVBF methods when $m = 500$ due to the discussion in the previous subsection.

CVBF$_K(\mathcal{S})$	γ	Scaled CVBF$_K(\mathcal{S})$	γ	CVBF$_K(\mathcal{D})$	γ	Freq. Test	γ
$m = 100$	0	$m = 100$	.001	$m = 100$	0	Mardia (Skew)	.999
$m = 200$	.026	$m = 200$	.717	$m = 200$	.086	Mardia (Kurt.)	.205
$m = 300$	.270	$m = 300$	.980	$m = 300$	.325	Royston	.554
$m = 400$	.611	$m = 400$	.994	$m = 400$	.538	Henze-Zirkler	.995

Table 4.3: The proportion of 1,000 randomly generated skew-normal random samples with $n = 1,000$ where each goodness-of-fit test correctly concludes against trivariate normality.

CVBF$_K(\mathcal{S})$	γ	Scaled CVBF$_K(\mathcal{S})$	γ	CVBF$_K(\mathcal{D})$	γ	Freq. Test	γ
$m = 100$	.005	$m = 100$	.249	$m = 100$	.223	Mardia (Skew)	.649
$m = 200$	.160	$m = 200$	.992	$m = 200$	.992	Mardia (Kurt.)	1
$m = 300$	.278	$m = 300$	1	$m = 300$	1	Royston	1
$m = 400$	.381	$m = 400$	1	$m = 400$	1	Henze-Zirkler	1

Table 4.4: The proportion of 1,000 randomly generated Laplace random samples with $n = 1,000$ where each goodness-of-fit test correctly concludes against trivariate normality.

The results in Tables 4.3 and 4.4 are quite illuminating for both the kernel CVBF methods and frequentist tests. First, consider those in Table 4.3 for skew-normal data. From previous simulations and discussions in Sections 4.3 and 4.4, we already knew the power of the unscaled kernel CVBF methods was rather poor when testing normality for skew-normal data. This is exactly what we find over the 1,000 random samples as the empirical powers for CVBF$_K(\mathcal{S})$ and CVBF$_K(\mathcal{D})$ never exceed .611 and .538, respectively for suitable training set sizes. The empirical power for the scaled CVBF$_K(\mathcal{S})$ method is

rather impressive when $m \geq 300$ and far superior to its kernel CVBF counterparts. In fact, $\gamma \geq .980$ for the scaled CVBF$_K(\mathcal{S})$ method is comparable to Mardia's test based on multivariate skewness ($\gamma = .999$) and the Henze-Zirkler test ($\gamma = .995$). The version of Mardia's test that uses multivariate kurtosis performs rather poorly with empirical power $\gamma = .205$ and Royston's multivariate Shapiro-Wilk test rejects multivariate normality in just over half the random samples.

Next, much like the skew-normal data, the results for the Laplace data in Table 4.4 resemble what we have seen in previous simulations. Both the scaled CVBF$_K(\mathcal{S})$ method and the CVBF$_K(\mathcal{D})$ method perform very similarly with at least 992 of the 1,000 samples favoring non-normality when $m \geq 200$. Since the scale parameters of each coordinate of the Laplace distribution are randomly chosen, the covariance matrix is not necessarily proportional to the identity matrix; however, it is a diagonal matrix. Therefore, the poor performance of the CVBF$_K(\mathcal{S})$ construction and the great performance of the CVBF$_K(\mathcal{D})$ method are not surprising. For the frequentist tests, the Henze-Zirkler test performs very well once again with all 1,000 samples rejecting normality at the $\alpha = .001$ level. For the heavy tailed Laplace data, Mardia's test based on multivariate kurtosis performs perfectly well, while the test based on skewness rejects normality in 65% of samples, which is opposite what we saw with the skew-normal data.

4.7.3 Conclusions

Overall, the results from this simulation illustrate why we prefer a Bayesian approach to goodness-of-fit testing. The Type I error rates for the kernel CVBF methods are far superior to the frequentist tests. A frequentist would have to set $\alpha < .001$ in order for the Type I error rates to agree with our Bayesian approach. For larger data sets, this significance level will need to tend to 0 since the kernel CVBF methods are consistent. Once we set the Type I error rates to be roughly the same for all tests, the Henze-Zirkler test

performed quite well, on par with the scaled $\text{CVBF}_K(\mathcal{S})$ approach for appropriate training set size. The other frequentist tests, both of Mardia's approaches and Royston's method, perform well for one alternative model, but not the other. Therefore, combining the excellent performance under both the null and alternative models, as well as the performance in the many previous simulations in this chapter, we can safely conclude that the scaled $\text{CVBF}_K(\mathcal{S})$ method is the superior kernel CVBF approach.

4.8 Curse of Dimensionality

In Section 3.4, we provided two different definitions for the curse of dimensionality, namely the increase in computational complexity and the "empty space phenomenon" that occurs as the data dimension increases. We also described the drastic impact that this curse has when nonparametrically estimating a multivariate density function, specifically, the fact that the typical multivariate kernel density estimate should not be used for data in more than 5 dimensions.

In this section, we will answer the questions posed in Section 3.5 relating to the effect the curse of dimensionality has on the multivariate kernel CVBF method. We will first get a sense of the applicability of the kernel CVBF method to higher dimensional data using a simulation for testing 10-dimensional normality. After discussing the potential pitfalls of testing goodness-of-fit of high dimensional data with the kernel CVBF method, we offer viable work-around solutions to the curse by means of dimension reduction techniques.

4.8.1 The Impact of the Curse of Dimensionality on Kernel CVBF Methods

If the kernel density estimate is only appropriate for estimating data in moderate dimensions, how well would it work in a 10-dimensional case? So far in this chapter, we have concluded that applying the scaled $\text{CVBF}_K(\mathcal{S})$ method is the preferred kernel CVBF approach as it outperforms and is far simpler than its kernel CVBF counterparts. The kernel estimate simply needs to pick up on major departures from the null model. Consider

testing normality once again, but this time in 10 dimensions. We draw 32 random samples of size $n = 5,000$ from the standard normal $(N(\mathbf{0}, \mathbf{I}_{10}))$, skew-normal $(SN(\mathbf{0}, \mathbf{I}_{10}, \mathbf{10}))$, and independent Laplace (each coordinate $L(0, 1)$) distributions. For each distribution, the scaled $\text{CVWE}_K(\mathcal{S})$ values are computed for training set sizes $m = 500, 1000, 1500, 2000,$ and 2500 using $N = 30$ random splits. The results are in Figure 4.8.

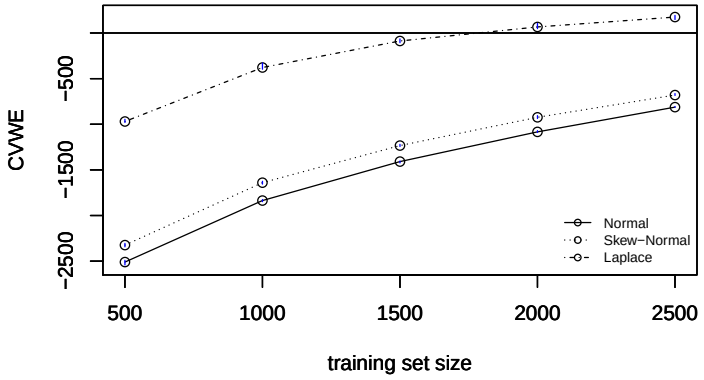

Figure 4.8: Testing 10-dimensional normality using the scaled $\text{CVBF}_K(\mathcal{S})$ method. The simulation consists of 32 samples from normal (solid), skew-normal (dotted), and Laplace (dotdashed) distributions, $N = 30$ random splits, and training set sizes $m = 500, 1000, 1500, 2000,$ and 2500.

Not surprisingly, the scaled $\text{CVBF}_K(\mathcal{S})$ method performs very well under the null model. We expect the parametric model to be favored when the null model is true compared to a nonparametric estimate. With only $n = 5,000$ observations, the scaled $\text{CVBF}_K(\mathcal{S})$ method concludes in favor of the normal model for skew-normal data. We saw earlier in this chapter (Figure 4.3) how transforming the data to have identity covariance matrix offered improved performance for the skew-normal model when $d < 5$. Now, for higher dimensional data, the scale transformation is no longer a suitable remedy for skew-normal

data. Maybe most impressively, with a training set size $m \geq 1750$, even in 10 dimensions, we conclude in favor of non-normality for the Laplace distribution. However, the major takeaway from Figure 4.8 is that while the scaled $\text{CVBF}_K(\mathcal{S})$ method can be used to test goodness-of-fit in more than moderate dimensions, the multivariate kernel density estimate is too inaccurate without massive data sets to have confidence in the conclusions.

The difficulty in the scaled $\text{CVBF}_K(\mathcal{S})$ method in successfully favoring the kernel model for skew-normal data in 10 dimensions is directly related to the second definition of the curse of dimensionality. Comparing Figures 4.8 and 4.7, for the same skew-normal distribution with the same parameters (save for dimension) and $n = 5,000$ the scaled $\text{CVBF}_K(\mathcal{S})$ method overwhelmingly favors non-normality when $d = 4$, but when $d = 10$, we would overwhelmingly favor normality. The following results are not shown in any figure, but in two dimensions, we only need $n = 500$ and $m \geq 100$ to reach the correct conclusion for skew-normal data. However, in order to favor the kernel model in three and four dimensions, the (n, m) pairs required are $(2000, 500)$ and $(5000, 2000)$, respectively. The increasing training set size reflects the "empty space phenomenon" since we need a far larger number of observations in order for the kernel model to detect the skewness and reject the normal model. Therefore, it is no surprise that with only $n = 5,000$ observations, the scaled $\text{CVBF}_K(\mathcal{S})$ method favors normality.

For other alternative models, the increased dimension does not have as drastic an effect on the number of observations required to reject normality. In fact, for the four-dimensional Laplace model the scaled $\text{CVBF}_K(\mathcal{S})$ method favors the kernel model with sample size $n = 500$ and training sample size $m = 150$. We can even conclude against normality in the 10-dimensional Laplace model in Figure 4.8 with $m = 1750$ for $n = 5000$. Certainly the curse of dimensionality impacts different densities with different amounts of severity. Unfortunately, we do not know this severity prior to analyzing a single data set from an unknown density function.

The effect of increasing data dimensions on computing the kernel CVBF is far less substantial, even minimal, provided that we only consider scalar bandwidth matrices. We have already mentioned previously in this chapter that the computation time increases dramatically for the diagonal and full bandwidth matrix classes. For bivariate data, compared to the $\text{CVBF}_K(\mathcal{S})$ method, computing the $\text{CVBF}_K(\mathcal{D})$ and $\text{CVBF}_K(\mathcal{F})$ values for 500 normal observations ($m = 100$, $N = 40$) takes roughly 4 (1 extra smoothing parameter) and 400 (2 extra smoothing parameters) times longer, respectively. These ratios are only exacerbated as d gets larger!

All that being said, if we restrict ourselves to the scaled $\text{CVBF}_K(\mathcal{S})$ method, we only incur a slight increase in computing time due to the matrix and vector calculations needed to evaluate the priors and likelihood functions. When $d = 2$, the computing time for a single random split of one data set ($n = 5000$, $m = 2000$) is 6.5 seconds compared to 22.7 seconds for the same scenario when $d = 10$. This increase is rather inconsequential in the long run and thus, the curse of dimensionality does not impact computation time significantly.

4.8.2 Dimension Reduction Techniques Applied to Kernel CVBF

Even though we may be able to apply the multivariate kernel CVBF method to data beyond moderate dimensions with some success as seen in the 10-dimensional example in Figure 4.8, unless the sample size is very large, we will undoubtedly favor the null model far too often in practice. In today's world of big data, it is common to encounter high-dimensional data sets in practice, so how can we utilize the scaled $\text{CVBF}_K(\mathcal{S})$ method to test goodness-of-fit for data when $d \geq 6$? In this subsection we provide two possible work-around solutions: test goodness-of-fit on all possible joint marginal distributions and perform goodness-of-fit tests after projecting to data into a lower dimensional space.

The most natural of the two solutions is to test the goodness-of-fit for all marginal

distributions of dimension $1 \leq d' < d$ from the original data set. This is natural because when $d \geq 3$, the only way we can visually examine the data is by looking at univariate and bivariate plots. Now, we simply extend this idea to the goodness-of-fit testing scenario.

The feasibility of this approach depends on the parametric null model being tested. For instance, we know that all d'-dimensional marginal distributions from a d-variate normal distribution are also normally distributed. More specifically, let $\mathbf{x} \sim N_d(\mu, \mathbf{\Sigma})$ and let $\mathbf{E}$ be a $d' \times d$ matrix where each row corresponds to a vector indexing a specific coordinate of $\mathbf{x}$. For instance, $\mathbf{E} = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right]$ refers to the bivariate marginal comprised of the first and third components of a trivariate normal distribution. Then, $\mathbf{Ex} \sim N_{d'}(\mathbf{E}\mu, \mathbf{E\Sigma E}')$. Similar results hold for other null models, particularly the multivariate t_3 distribution (Nadarajah and Kotz, 2005) and the multivariate skew-normal distribution (Azzalini and Capitanio, 1998).

In each of these distributions where the marginal distributions are members of the same family as the full data, assuming that $\binom{d}{d'}$ is not prohibitively large, we can test goodness-of-fit for each of the possible d'-dimensional marginal distributions. However, it is not necessarily true that if all possible d'-dimensional marginal distributions are from the null model, then the d-dimensional data are also from the parametric model. By testing goodness-of-fit on the marginal distributions, we can only show that the original data set *does not* follow the null model if at least one marginal distribution has a kernel CVBF value that favors the alternative model. Unfortunately, we would not be able to formally conclude in favor of the null model. However, much like a frequentist goodness-of-fit test, if all of the marginals were deemed to follow the null model by kernel CVBF, then we could argue that the parametric model is plausible.

Take the ever problematic skew-normal model ($SN(\xi = \mathbf{0}, \mathbf{\Omega} = \mathbf{I}_4, \alpha = \mathbf{10})$) as an example. For a single data set of $n = 1,000$ observations, taking a training set size of $m = 400$ produced a scaled $\text{CVWE}_K(\mathcal{S})$ value of 0.88 over $N = 30$ random splits.

According to the scale of Kass and Raftery (1995), $0.88 < \log(3)$, and thus we could not conclude definitively in favor of the kernel model. Suppose that instead, we tested each of the six bivariate marginal distributions for normality for this data set. For the original data $\boldsymbol{X}$, denote the lth coordinate vector to be $\boldsymbol{X}_{\cdot l}$. The scaled CVWE$_K(\mathcal{S})$ values for the six bivariate marginals $(\boldsymbol{X}_{\cdot i}, \boldsymbol{X}_{\cdot j})$ for $i \neq j$ are given in Table 4.6 using the same training set size and number of splits. Note that for each bivariate distribution, we re-scale the data using only the coordinates being tested.

$(\boldsymbol{X}_{\cdot i}, \boldsymbol{X}_{\cdot j})$	$\boldsymbol{X}_{\cdot 2}$	$\boldsymbol{X}_{\cdot 3}$	$\boldsymbol{X}_{\cdot 4}$
$\boldsymbol{X}_{\cdot 1}$	-5.22	3.31	-1.33
$\boldsymbol{X}_{\cdot 2}$		-5.00	-0.11
$\boldsymbol{X}_{\cdot 3}$			2.71

Table 4.5: Testing four-dimensional normality of $n = 1,000$ $SN(\xi = \mathbf{0}, \boldsymbol{\Omega} = \mathbf{I}_4, \alpha = \mathbf{10})$ observations using the scaled CVWE$_K(\mathcal{S})$ values from the six two-dimensional marginal distributions ($m = 400$, $N = 30$)

Of the six bivariate marginal distributions, the scaled CVWE$_K(\mathcal{S})$ values for $(\boldsymbol{X}_{\cdot 1}, \boldsymbol{X}_{\cdot 3})$ and $(\boldsymbol{X}_{\cdot 3}, \boldsymbol{X}_{\cdot 4})$ indicate strong and positive evidence against normality, respectively. Based on these conclusions, we would doubt the normality of the original four-dimensional data set. One word of caution when using the skew-normal model in practice. The skew-normal distribution we consider in the simulation has parameters $\xi = \mathbf{0}$, $\boldsymbol{\Omega} = \mathbf{I}$, and $\alpha = \mathbf{10}$. From Proposition 2 of Azzalini and Capitanio (1998), it is true that the bivariate marginals are indeed skew-normal distributions as well; however, the skew parameter is attenuated. Ignoring the scale transformation, each of the bivariate marginal distributions

has parameters $\xi' = \mathbf{0}$, $\mathbf{\Omega}' = \mathbf{I}$, and $\alpha' = \frac{\mathbf{10}_2}{(1+\mathbf{10}_2^T\mathbf{10}_2)^{1/2}} = .705$. This attenuation still exists after transforming the data to have identity covariance matrix and is exacerbated as the difference $d - d'$ increases. Similar phenomena may exist for other distributions, so even though this approach of testing the d'-dimensional marginal distributions works well in this example, it is not foolproof.

Another solution to the large d problem is to apply dimension reduction techniques to the full data set and perform a goodness-of-fit test using kernel CVBF on the reduced dimension space. This is a common two-stage approach to estimating a density function for high-dimensional data using kernel density estimates (Biau and Mas, 2010). As Scott (1992) points out, the underlying structure of d-dimensional data is often d''-dimensional where $d'' << d$. He claims that, in practice, data of any dimension can be reduced down to a four- or five-dimensional structure. This implies that the kernel CVBF approach can be applied to any goodness-of-fit problem once a dimension reduction procedure is applied to the original data. There are numerous dimension reduction techniques in the literature (see Fodor (2002) for a list), including principal components analysis, factor analysis, and projection pursuit. We only consider principal components at this time because it is the simplest to implement and the resulting components are linear combinations of the original data vectors. The goal is to find the underlying structure that explains more than 95% of the variation in the original data with fewest number of dimensions d''. Then, apply the scaled CVBF($\mathcal{S}$) method to the d''-dimensional principal component transformed data. The goodness-of-fit conclusions for the original data based on the kernel CVBF results for the reduced data will again be subject to restrictions depending on the properties of the null parametric model. This approach would work well for the t_3, normal, and skew-normal models since any linear combination of the d coordinates will follow a distribution within the same family. However, once again, we can only conclude against and never in favor of the parametric model. Also, this approach can be problematic in that we are not assured

of reducing the data dimension to d'' that is small enough for the kernel CVBF method to behave appropriately.

Both of these approaches allow us to apply the kernel CVBF method to test distributional goodness-of-fit of any parametric model when the data dimension gets large. However, we do have to pay a penalty for making the test simpler through dimension reduction. That penalty comes in the form of the possible conclusions we can make regarding the full data. No longer can we simply conclude in favor of either the null or alternative models. Even in the reduced dimension cases, we can still favor the alternative model, but regarding the null model, at best we can say it is a plausible data generating model. These two conclusions have the same feeling as "rejecting" and "failing to reject" the null model in a frequentist goodness-of-fit test.

4.9 Data Analysis

From 1999-2013, the state of California assessed the academic performance of schools (Elementary, Middle, and High School) using standardized tests in accordance with the Public Schools Accountability Act of 1999. Based on these tests, schools are ranked according to their Academic Performance Index (API) which can range from 200 to 1,000. The higher the API, the better students performed on the test, but the goal is to have all schools above 800. For more information about API and full reports/data, see the California Department of Education webpage (www.cde.ca.gov/ta/ac/ap/).

In order to see how to apply the scaled $\text{CVBF}_K(\mathcal{S})$ method to test goodness-of-fit in practice, consider testing bivariate normality for the following data taken from the *survey* package in R (Lumley, 2017). There are 757 school districts in California with at least one school having more than 100 enrolled students and 570 of these 757 districts have two or more such schools. Two schools are randomly selected from each of the 570 districts and a bivariate kernel density estimate of the API scores from 2000 is plotted in the contour

plot in the left panel of Figure **??**. From the contour plot, the distribution of API values appears to be roughly normally distributed with some slight bimodality. To compare the API data to normal data, the right panel of Figure **??** contains the contour plot of 570 randomly sampled bivariate normal observations with location and scale parameters set to the parameter estimates from the API data. Certainly, the two distributions are very similar, with the exception of the extra mode.

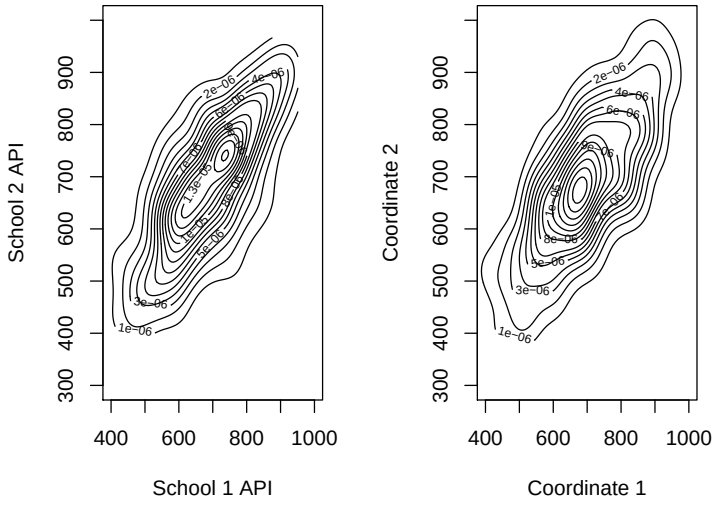

Figure 4.9: *Left Panel*: Contour plot displaying the bivariate distribution of API scores from the year 2000 for two schools chosen from 570 districts in California. *Right Panel*: Contour plot of 570 observations from a $N_2\left(\hat{\mu}, \hat{\mathbf{\Sigma}}\right)$ distribution based on the sample estimates from the API data.

To carry out the goodness-of-fit test, we follow the steps of calibration in Subsection 4.5.1. We first transform the API data to have identity sample covariance. Therefore, for each of the $n = 570$ observed API values (X_i), we compute $Y_i = \left[\begin{smallmatrix} 13738 & 10663 \\ 10663 & 14396 \end{smallmatrix}\right]^{-1/2} X_i$. Next, we compute the scaled $\text{CVWE}_K(\mathcal{S})$ value for training set sizes $m = \{30, 31, \ldots, 284, 285\}$

and $N = 52$ random splits. The resulting $(m, \text{CVWE}_{m,52})$ pairs are plotted in Figure 4.10 along with curves for the respective first and third quartiles. Ignoring the curve for the null data for now, what makes this such an interesting example is the fact that for $m \leq 75$, we would find positive evidence in favor of the normal model, but when $m \geq 100$ we would conclude positively against normality. This illustrates the importance of comparing the observed kernel CVWE values to those from data sampled from the null model.

For each of 25 independent random samples of size $n = 570$ from the estimated bivariate normal model, we compute $\text{CVWE}_{m,20}$ for training set sizes $m = 30, 34, \ldots, 281, 285$ using the scaled $\text{CVBF}_K(\mathcal{S})$ method. The black dashed curve in Figure 4.10 represents the median $\text{CVWE}_{m,20}$ value across the 25 null samples. Similarly, the gray dashed curves below and above the median curve represent the first and third quartiles of the 25 CVWE values, respectively. The median CVWE curve is always negative and only for training set sizes larger than $m = 260$ does the third quartile curve take a nonnegative value. By simply comparing the observed and null CVWE curves, bivariate normality is unlikely for the API data since the curves differ so greatly.

In order to choose a training sample size we continue with the calibration steps. We can see that using any training set size $m \geq 150$ produces essentially the same conclusions, with a CVWE value more than 10. According to Kass and Raftery (1995), a CVWE value greater than 5 represents very strong evidence against the null. Remember that we want to choose m such that the observed CVWE curve is (nearly) maximized, while at the same time the null CVWE curve is as small as possible. The strict maximum of the observed CVWE curve occurs for $m = 277$ ($\text{CVWE}_{277,52} = 17.64$), but at this training set size, the null CVWE curve is very close to 0. To follow the calibration rules, instead of taking $m = 277$, consider $m = 170$ (30% / 70% split) such that $\text{CVWE}_{170,52} = 14.95$ for the API data. Not much is lost in terms of the amount of evidence in favor of the kernel model, but much is gained under the null model since the CVWE value is -8.92. Therefore, we have

very strong evidence that the API data are *not* normally distributed.

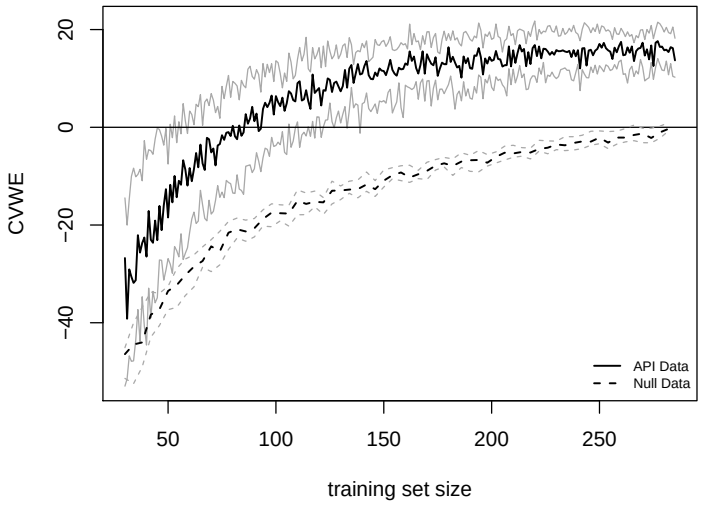

Figure 4.10: Scaled CVWE$_K(\mathcal{S})$ curves for the observed API data based on $N = 52$ random splits and bivariate normal data based on $N = 20$ splits for training set sizes $m = \{30, 31, \ldots, 284, 285\}$.

4.10 Application to Random Effects Models

One interesting application of the kernel CVBF method is in testing the distributional assumptions in random effects modelling. Let $X_1, X_2, \ldots, X_p$ comprise a random sample with $X_i = (X_{i1}, \ldots, X_{in}) \in \mathbb{R}^n$ and consider the simple random effects model given by:

$$X_{ij} = \mu_i + \epsilon_{ij}, \qquad j = 1, \ldots, n, \; i = 1, \ldots, p. \tag{4.15}$$

The typical assumptions for this model include:

- $\mu_1, \mu_2, \ldots, \mu_p \overset{iid}{\sim} N(\mu, \sigma_\mu^2)$.

- $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\epsilon^2) \quad \forall\, i, j.$
- μ and ϵ are independent of each other.
- $n \geq 2$.

Under these assumptions, the underlying distributions of the parameters μ_i and ϵ_{ij} follow a Gaussian model, which is by far the most common. If these distributional assumptions appear reasonable for the data at hand, then the subsequent analyses seek to determine if the two variance terms, σ_μ^2 and σ_ϵ^2 significantly differ from 0. Our goal here is to apply the scaled CVBF$_K(\mathcal{S})$ method to test the fit of the Gaussian model.

When $n = 2$, Reiersøl (1950) showed that the distribution functions F_μ and F_ϵ of μ_i and ϵ_{ij}, respectively, are completely determined by the joint distribution of (X_{i1}, X_{i2}) provided the characteristic functions of F_μ and F_ϵ meet mild regularity conditions. This result naturally extends to $n \geq 2$ using properties of the normal distribution, so we would simply need to verify that $(X_{i1}, \ldots, X_{in}) \sim N_n(\mu, \mathbf{\Sigma})$. We should point out that we could simply test univariate normality of the X_{ij}'s using the Hart and Choi (2016) method citing Cramér's Normal Decomposition Theorem. However, we will still apply the multivariate kernel CVBF method since for other parametric null models the marginal of X_{ij} does not determine the distributions of μ_i and ϵ_{ij}.

4.10.1 Formulation of the Null and Alternative Marginal Likelihoods

Under the null Gaussian model, define the variance of X_{ij} to be $\sigma^2 = \sigma_\mu^2 + \sigma_\epsilon^2$ and let $\rho = \sigma_\mu^2/\sigma^2$. Thus, we can parameterize the null model as $N_n\Big(\mu_n, \sigma^2[(1-\rho)\mathbf{I}_n + \rho\mathbf{J}_n]\Big)$. Based on this model, $X_{i1}, \ldots, X_{in}$ are exchangeable, implying that all n'-dimensional $(n' < n)$ marginal distributions are the same. Therefore, the major modification to the kernel CVBF methods when applied to random effects models is that we should force the kernel model to have the same marginal distributions of any dimension less than n.

Once the data are randomly split, the simple fix is to augment the training data using a Latin square such that each column of the augmented training data contains the same nm observations. Certainly, other augmentation schemes can be implemented, such as permuting all columns, but the computations become too unwieldy, even in moderate dimensions. We still transform the data to have identity sample covariance prior to computing the $\text{CVWE}_K(\mathcal{S})$ values; however, the augmented training data will no longer have identity sample covariance. That being said, the difference from identity is small enough to not be of any practical consequence. So computation of the alternative marginal likelihood remains the same as the scaled $\text{CVBF}_K(\mathcal{S})$ method provided we augment the training data set appropriately.

As for the null marginal likelihood, we can write the n-dimensional normal likelihood function as

$$L(\mathbf{X}|\mu,\sigma^2,\rho) = (2\pi)^{-pn/2}[\sigma^{2n}(1-\rho)^{(n-1)}(1+(n-1)\rho)]^{-p/2} \\ \times \exp\Big[-\frac{np\hat{\sigma}^2[1+\rho(n-2)+\rho\hat{\rho}(1-n)]+np(1-\rho)(\mu-\hat{\mu})^2}{2\sigma^2(1-\rho)(1+(n-1)\rho)}\Big],$$

where we have used the facts that

- $|\sigma^2[(1-\rho)\mathbf{I}_n+\rho\mathbf{J}_n]| = \sigma^{2n}(1-\rho)^{n-1}(1+(n-1)\rho)$,
- $\mathbf{S} = adjoint(\mathbf{\Sigma})$ where $S_{jj} = \sigma^{2(n-1)}(1-\rho)^{n-2}(1+(n-2)\rho)$, and $S_{jj'} = S_{j'j} = -\sigma^{2(n-1)}\rho(\rho-1)^{n-2},\ j \neq j'$, and
- $\mathbf{\Sigma}^{-1} = |\mathbf{\Sigma}|^{-1}\mathbf{S}$.

Consider the following UIR prior distributions for μ, σ^2, and ρ:

- $\pi(\mu|\sigma^2,\rho) = (2\pi)^{-1/2}n^{1/2}[\sigma^2(1+(n-1)\rho)]^{-1/2}\exp\Big[-\frac{1}{2}\frac{n(\mu-\hat{\mu})^2}{\sigma^2(1+(n-1)\rho)}\Big]$,
- $\pi(\sigma^2|\rho) = \frac{n\hat{\sigma}^2[1+\rho(n-2)+\rho\hat{\rho}(1-n)]}{2(1-\rho)(1+(n-1)\rho)}(\sigma^2)^{-2}\exp\Big[-\frac{1}{\sigma^2}\frac{n\hat{\sigma}^2[1+\rho(n-2)+\rho\hat{\rho}(1-n)]}{2(1-\rho)(1+(n-1)\rho)}\Big]$, and

- $\pi(\rho) = 1_{[0,1]}(\rho)$,

where the parameter estimates $\hat{\mu}$, $\hat{\sigma}^2$, and $\hat{\rho}$ are given by

- $\hat{\mu} = \frac{1}{np}\sum_{i=1}^{p}\sum_{j=1}^{n} X_{ij}$,
- $\hat{\sigma}^2 = \frac{1}{np}\sum_{j=1}^{n}\sum_{i=1}^{p}(X_{ij} - \hat{\mu})^2$, and
- $\hat{\rho} = \frac{2}{n(n-1)\hat{\sigma}^2 p}\sum_{i=1}^{p}\sum_{k=1}^{n-1}\sum_{j=k+1}^{n}(X_{ik} - \hat{\mu})(X_{ij} - \hat{\mu})$.

The resulting marginal likelihood function is as follows:

$$\begin{aligned} M_0(\mathbf{X}) &= \int_0^1 \int_0^\infty \int_{-\infty}^\infty L(\mathbf{X}|\mu, \sigma^2, \rho)\pi(\mu|\sigma^2, \rho)\pi(\sigma^2|\rho)\pi(\rho)d\mu d\sigma^2 d\rho \\ &= \pi^{-\frac{pn}{2}}\Gamma\Big(\frac{np+2}{2}\Big)(p+1)^{-\frac{(np+3)}{2}}(n\hat{\sigma}^2)^{-\frac{pn}{2}} \\ &\times \int_0^1 (1-\rho)^{-\frac{p(n-1)}{2}}(1+(n-1)\rho)^{-\frac{p}{2}}\Big[\frac{1+\rho(n-2)+\rho\hat{\rho}(1-n)}{(1-\rho)(1+(n-1)\rho)}\Big]^{-\frac{pn}{2}} d\rho. \end{aligned} \tag{4.16}$$

4.10.2 Random Effects Model Simulation ($n = 2$)

In order to verify that the scaled CVBF$_K(\mathcal{S})$ method can adequately test the appropriateness of the Gaussian model in a simple random effects model (4.15), consider the following small simulation. Using the marginal likelihoods given in Subsection 4.10.1, consider $n = 2$ and $p = 1,000$. Let $\mu = 0$ so that $X_{ij} \sim$ Laplace$(0, 1)$ or $X_{ij} \sim N(0, 1)$. For each data set, we draw 25 independent random samples and compute the scaled CVWE$_K(\mathcal{S})$ values using training set sizes $m = 100, 200, 300, 400,$ and 500 as well as $N = 30$ random splits of the data. The results are given in Figure 4.11. Certainly, the scaled CVBF$_K(\mathcal{S})$ method correctly concludes in favor of the null model for normal data and in favor of the kernel model for Laplace data regardless of training set size. Thus, the scaled CVBF$_K(\mathcal{S})$ method is reasonably well-suited to checking the Gaussian model assumption in random effects models.

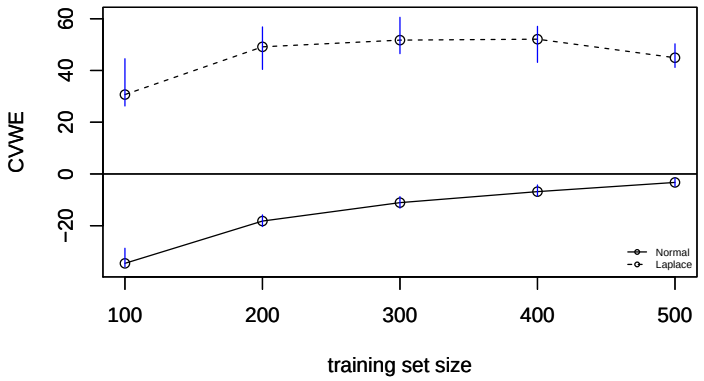

Figure 4.11: Verifying the applicability of the scaled $\text{CVBF}_K(\mathcal{S})$ method to check the Gaussian model assumption in a simple random effects model. For 25 samples, either $X_{ij} \sim L(0,1)$ (dashed line) or $X_{ij} \sim N(0,1)$ (solid line), of size $p = 1000$ and dimension $n = 2$, the $\text{CVWE}_K(\mathcal{S})$ values are computed using $N = 30$ and $m = 100, 200, \ldots, 500$.

4.10.3 Real Data Example: Gene Expression Levels in Rats

Davidson et al. (2004) conducted a study to explore the effect of n-3 polyunsaturated fatty acids on colon cancer tumors in ninety rats using gene expression analysis. The data we consider here is a subset of $n = 5$ rats, each with expression levels for $p = 8,038$ genes. The expression levels for each rat have been demeaned such that the average expression level for all 8,038 genes is 0 for each rat. The distributions of all expression levels for each rat are plotted in Figure 4.12.

Suppose, for the purposes of this example, the researchers were interested in determining how much of the variability in rat gene expression levels is due to the overall gene effect σ_μ^2 and the gene effect within each rat σ_ϵ^2. Based on the distributions in Figure 4.12, there appears to be a rat effect since there are two distinct groups of rats($\{1,2\}$, $\{3,4,5\}$). To answer these questions, we would use the model in (4.15), which means we need to assume the Gaussian model. We will use the scaled $\text{CVBF}_K(\mathcal{S})$ method to address this

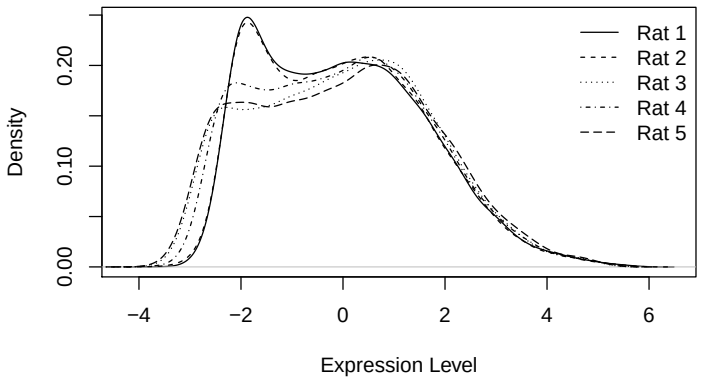

Figure 4.12: The estimated distribution of gene expression levels for each of $n = 5$ rats and $p = 8,038$ genes from the colon cancer study conducted by Davidson et al. (2004).

assumption.

Using the entire data set and computing the scaled $\text{CVWE}_K(\mathcal{S})$ value for $N = 30$ random splits and training sample size $m = 1,606$, we find that $\text{CVWE}_{1606,30} = 19,914.01$, which represents astronomical evidence against the Gaussian model. We do not consider more training set sizes because regardless of m, there is no debate about the inappropriateness of the Gaussian model.

In order to compute the CVWE value over the entire data set, the corresponding run time was 2,352 seconds. Could we use a *Divide and Conquer* scheme to make this even faster? Consider choosing $w = 10$ such that 2 subgroups have $p = 803$ and 8 subgroups have $p = 804$ genes. Using $N = 30$ and a 30% / 70% split into training and validation sets in each subgroup, we find an overall CVWE value of $1,641.16$ in just 202 seconds. We use a larger training set proportion in the subgroups since we are estimating a five-dimensional distribution with only 800 observations in each subgroup. Keep in mind that as n gets larger, choosing a larger training sample proportion will cause a more dramatic increase

in computation time due to the augmentation. However, we still find an overwhelming amount of evidence against the Gaussian model but in 8.5% of the time.

Since we cannot visualize five-dimensional data to see if the gene expression levels follow a multivariate normal distribution, we can combine the dimension reduction idea of Subsection 4.8.3 with the *Divide and Conquer* scheme to quickly compute the CVWE values on the 10 bivariate marginal distributions. First, if we look at the bivariate marginal distributions in Figure 4.13 the "dagger" shape and/or slight curvature of the contour plots give us reason to doubt bivariate normality.

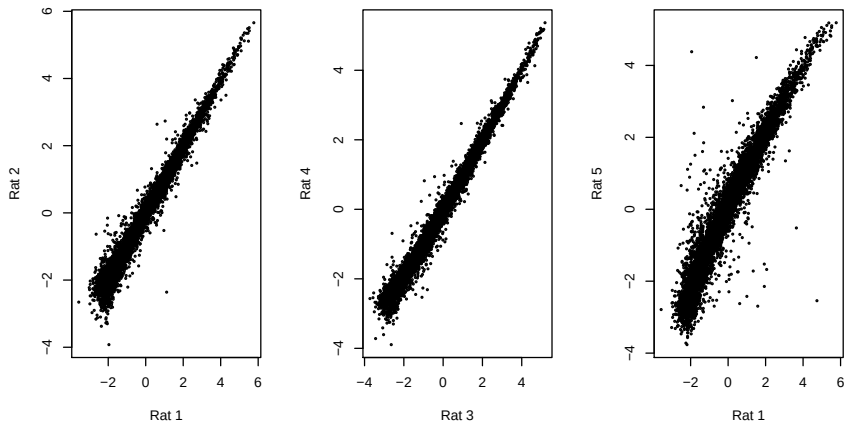

Figure 4.13: Bivariate scatterplots of gene expression levels for the pairs of rats: (1,2) (*left panel*), (3,4) (*middle panel*), and (1,5) (*right panel*).

For the *Divide and Conquer* scheme, we use the same setup as before for the full data set with $w = 10$ partitions and $N = 30$ random splits, but now consider a 20% / 80% allotment to the training and validation sets. The resulting CVWE values for each of the 10 bivariate distributions are provided in Table 4.6. The corresponding run time for each CVWE value was roughly 56 seconds and the smallest CVWE value between any two rats

is 28.32, for rats 1 and 5. Once again, on the scale of Kass and Raftery (1995), we have overwhelming evidence against bivariate normality for each of the 10 marginals, hence the five-dimensional data are also non-normal.

	Rat 2	Rat 3	Rat 4	Rat 5
Rat 1	68.57	35.73	62.75	44.49
Rat 2		28.32	63.87	63.93
Rat 3			47.84	40.10
Rat 4				33.01

Table 4.6: Scaled $\text{CVWE}_K(\mathcal{S})$ values for testing bivariate normality for the 10 bivariate marginal distributions for the $n = 5$ rats using a *Divide and Conquer* scheme with $w = 10$, $N = 30$, and 20% / 80% split.

Clearly, whether we consider the full five-dimensional gene expression data set or the bivariate marginal distributions, there is overwhelming evidence against the Gaussian model for these rat data. The real difference in these approaches is the computation time, which can be greatly reduced using a *Divide and Conquer* scheme and/or a dimension reduction approach. Regardless, the assumptions for the random effects model are not met and another analysis must be considered as the findings from the standard linear mixed model method might be invalid.

4.11 Summary and Conclusions

Over the course of this chapter, we have shown that the kernel CVBF methods of Hart and Choi (2016) can be extended to test distributional goodness-of-fit in the multivariate setting. For completeness, we developed separate constructions for each of the three

bandwidth matrix classes. However, the scaled CVBF$_K(\mathcal{S})$ construction proved to be superior in terms of performance and simplicity. This is a surprising finding since typically, the simplest method rarely outperforms more sophisticated approaches. Not only does it outperform the CVBF$_K(\mathcal{D})$ and CVBF$_K(\mathcal{F})$ methods in terms of finding evidence for or against the null model, but it also takes fractions of the time to compute.

We also looked at some important properties of the goodness-of-fit test based on the scaled CVBF$_K(\mathcal{S})$ construction. Provided that we re-scale the data to have identity covariance matrix, the kernel CVBF method is location-scale invariant, which is very important in a goodness-of-fit test. More importantly, we argued both mathematically and empirically that the CVBF$_K(\mathcal{S})$ method is consistent at an exponential rate under both hypotheses. The consistency of the CVBF$_K(\mathcal{S})$ implies that the probability of making a Type I or Type II error tends to 0 for increasing sample size. This makes the scaled CVBF$_K(\mathcal{S})$ method superior to common frequentist tests for normality since the Type I error rate remains constant as sample size increases. In order for the frequentist and kernel CVBF methods to agree in terms of Type I error rate, the significance level in the frequentist test must tend to 0!

The curse of dimensionality is still a concern for the kernel CVBF method. Whether we talk about increasing d and/or n the computation and feasibility of testing goodness-of-fit using the kernel CVBF method becomes difficult and risky. By restricting ourselves to only using the CVBF$_K(\mathcal{S})$ approach, increasing d does not greatly impact computation time. However, as many authors have discussed, the standard multivariate kernel density estimate is very poor beyond moderate dimensions. This is due to the "empty space phenomenon" and can only be remedied by considering extremely large data sets. This is a catch-22 for the kernel CVBF method because computation time increases quadratically with increasing sample size. A potential remedy to the large d problem is to reduce the dimension by considering subset marginal distributions or statistical dimension reduction

techniques. The solution to large n is to use a *Divide and Conquer* approach, which is made possible thanks to consistency. Both of these remedies can be applied simultaneously to a large n, large d data set, though we should proceed with extreme caution in this scenario. However, this means we can essentially make a recommendation regarding the goodness-of-fit of any multivariate model.

The real data examples illustrate how we might assess goodness-of-fit using the scaled $\text{CVBF}_K(\mathcal{S})$ method in practice. More specifically, the API data example walks through the steps of calibration to show the importance of selecting the training sample size appropriately. The gene expression data example explores how the scaled $\text{CVBF}_K(\mathcal{S})$ method can be modified and applied to check the model assumptions in a simple random effects model. We also saw how to implement the dimension reduction and *Divide and Conquer* ideas to a real data set using the rat data.

Overall, the scaled $\text{CVBF}_K(\mathcal{S})$ method is a very simple, intuitive, and computationally efficient Bayesian nonparametric approach to testing multivariate goodness-of-fit.

5. COMPARING TWO PARAMETRIC MODELS USING CVBF

5.1 Introduction

In Chapters 1 to 4, we were interested in comparing a parametric null model to a nonparametric alternative model using Bayes factors. While both the null and alternative models were required to be well-defined, no assumptions regarding the functional form of the alternative model were made. This is typical in the goodness-of-fit testing problem because we want to compare the parametric model to all other possible models. The question we are often asking in the hypotheses is, "Can we reasonably conclude that the data were sampled from the parametric model? If the answer to this question is "yes" based on the Bayes factor, then we proceed assuming the parametric model generated the data. If the answer is "no", we simply know that at least one nonparametric model is a better representation of the data and further exploration of the data is required. Certainly, this is an unsatisfactory conclusion, however it is simply the nature of the nonparametric goodness-of-fit test.

Now suppose that we have two competing parametric models which we would like to compare. This is the quintessential Bayes factor problem dating at least to Jeffreys (1961), which we described briefly in Subsection 1.1.2. Let M_1 represent the null parametric model with parameter space Λ and M_2 represent the alternative parametric model with parameter space Θ (both parameter spaces are subsets of multidimensional Euclidean spaces). Suppose we observe a data vector $\mathbf{x}$. When M_1 is the true model, the likelihood function for $\mathbf{x}$ is $L_1(\mathbf{x}|\lambda)$ and when M_2 is true, $L_2(\mathbf{x}|\theta)$ is the corresponding likelihood function. For prior distributions $\pi_1(\lambda)$ and $\pi_2(\theta)$, the corresponding marginal likelihoods are

$$m_1(\mathbf{x}) = \int_\Lambda \pi_1(\lambda) L_1(\mathbf{x}|\lambda)\, d\lambda \quad \text{and} \quad m_2(\mathbf{x}) = \int_\Theta \pi_2(\theta) L_2(\mathbf{x}|\lambda)\, d\theta.$$

Using these marginal likelihood, the Bayes factor for comparing M_1 and M_2 is given by

$$BF = \frac{m_2(\mathbf{x})}{m_1(\mathbf{x})},$$

the ratio of the posterior odds to the prior odds in favor of the alternative model.

In some cases, the marginal likelihoods are analytically tractable due to the choice of prior distribution ((semi-) conjugate prior) or the form of the likelihood (Gaussian). We saw in Chapter 4, that when testing multivariate normality, closed form solutions were easily attainable for the null marginal likelihood. However, in the majority of instances, the marginal likelihoods require some form of numerical integration (see Subsection 4.2.4) to evaluate. When the dimension of the parameter space becomes moderately large and/or the number of constraints imposed on the parameters becomes too complex, these integration techniques become time consuming and potentially impossible. The Laplace approximation in (4.10) is ideally suited for computing the marginals due to the approximate normality of the posterior distribution; however, we must numerically compute the posterior mode and the observed Hessian matrix of the posterior distribution. As we will see, there are instances where the Hessian matrix (or its inverse) do not exist for certain parameter values, rendering the Laplace approximation useless.

Another less than ideal aspect of Bayes factors is their behavior when the two models are nested. When the smaller model is true, both models are so similar that the Bayes factor has more difficulty choosing the correct model. Thus, the Bayes factor is typically consistent at a "power-of-n" rate (for sample size n). This is much slower than the exponential rate achieved when the larger model is true, where the Bayes factor more easily favors the larger model. This scenario of consistency for nested models was the driving force behind the non-local priors solution of Johnson and Rossell (2010).

In this chapter, we propose the parametric CVBF (CVBF$_P$) method that (a) uses Bayes

factors, (b) does not require prior distributions for the parameters in each model, (c) is computationally simple, and (d) is Bayes consistent at an exponential rate for both nested and non-nested models regardless of which model is true. The main crux of the approach is data splitting, which made the kernel CVBF methods of Chapters 2 and 4 possible. In the parametric CVBF method, we compute maximum likelihood estimates (MLEs) on a subset of the data and then compute a Bayes factor, which turns out to be a likelihood ratio, on the remainder of the data. So the difference in the two CVBF approaches is the training set was used to make the kernel model well-defined in the kernel CVBF method, but in the parametric CVBF method, both parametric models are determined using the MLEs from the training data. The simplicity in (c) stems from no longer needing to integrate or compute Hessian matrices to evaluate the marginal likelihoods.

The idea of data splitting has been used in Bayesian statistics by other researchers in contexts other than the previously described CVBF methods. For instance, Rust and Schmittlein (1985) apply a Bayesian cross-validated likelihood using leave-one-out cross-validation for model selection using posterior probabilities instead of Bayes factors. Another common use involves taking noninformative, improper priors and making them proper so they are suitable for model comparison in Bayes factors. Some examples include: intrinsic Bayes factors (Berger and Perrichi 1996, 2004), partial Bayes factors (O'Hagan, 1991), pseudo Bayes factors (Geisser and Eddy, 1979), and posterior Bayes factors (Aitkin, 1991). Each of these methods form a proper prior on the training set of various size (as small as a single observation and as large as the whole data set) which is then used as the prior for the entire data set.

The remainder of this chapter contains the following sections. First, we introduce the parametric CVBF (CVBF_P) methodology in Section 5.2. In Section 5.3, we discuss the necessary consistency results in both nested and non-nested cases and how they relate to other frequentist and Bayesian tests. Numerous different simulation studies and a real data

example are provided in Sections 5.4 and 5.5 to show the applicability and performance of the parametric CVBF approach for model comparison. Lastly, some overall conclusions are given in Section 5.6.

5.2 CVBF$_P$ Methodology

Before we describe the overall methodology of the parametric CVBF approach, we first examine the form of the Bayes factor when testing two simple hypotheses. Suppose we have observed data $\mathbf{x}$ and we want to compare to densities f_1 and f_2, both of which are fully specified densities for $\mathbf{x}$. If p and $1-p$ are the prior probabilities for f_1 and f_2, respectively, the posterior probabilities of f_1 and f_2 are

$$P(f_1|\mathbf{x}) = \frac{pf_1(\mathbf{x})}{pf_1(\mathbf{x}) + (1-p)f_2(\mathbf{x})}$$

and

$$P(f_2|\mathbf{x}) = \frac{(1-p)f_2(\mathbf{x})}{pf_1(\mathbf{x}) + (1-p)f_2(\mathbf{x})}.$$

Taking the ratio of the posterior probabilities, it follows that

$$\frac{P(f_2|\mathbf{x})}{P(f_1|\mathbf{x})} = \frac{f_2(\mathbf{x})}{f_1(\mathbf{x})}\frac{1-p}{p},$$

which means the Bayes factor is $f_2(\mathbf{x})/f_1(\mathbf{x})$ (Kass and Raftery, 1995). The important point here is that the Bayes factor *only* depends on the likelihood ratio and is free of prior probabilities on the parameters.

Now assume that $\mathbf{X} = (X_1, \ldots, X_n)$, $X_i \in \mathbb{R}^d$, are a random sample from some density function f_0. Let $M_1 = \{g(\cdot|\lambda) : \lambda \in \Lambda\}$ and $M_2 = \{f(\cdot|\theta) : \theta \in \Theta\}$ be parametric models for f_0, where Λ and Θ are subsets of q and p dimensional Euclidean spaces, respectively. Since we do not require prior distributions for λ and θ, the Bayes

factor is completely determined by the likelihood functions, which when computed on the whole data set are given by, $L_1(\lambda) = \prod_{i=1}^{n} g(X_i|\lambda)$ and $L_2(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$.

In order for this method to work, we need to randomly split the data into a training set $\mathbf{X}^T = (X_1, \ldots, X_m)$ and a validation set $\mathbf{X}^V = (X_{m+1}, \ldots, X_n)$. Note that the training and validation sets are mutually exclusive and exhaustive. Define the MLEs of λ and θ to be $\hat{\lambda}_m$ and $\hat{\theta}_m$, respectively, computed on the training data. Using these MLEs, $f(\cdot|\hat{\theta}_m)$ and $g(\cdot|\hat{\lambda}_m)$ are fully specified, simple models for the underlying distribution of X_i. Thus the Bayes factor is the likelihood ratio

$$\text{BF}_m(\mathbf{X}^T, \mathbf{X}^V) = \frac{\prod_{j=m+1}^{n} f(X_j|\hat{\theta}_m)}{\prod_{j=m+1}^{n} g(X_j|\hat{\lambda}_m)}, \tag{5.1}$$

computed from the validation data. Notice the slight, but very important difference between the Bayes factor in (5.1) and the classical likelihood ratio statistic. The standard likelihood ratio statistic for this test (provided the two models are nested) is given by $\frac{L_2(\hat{\theta}_n)}{L_1(\hat{\lambda}_n)}$, where $\hat{\theta}_n$ and $\hat{\lambda}_n$ are the MLEs from all n observations (i.e. the likelihood ratio *and* MLEs come from the same data). However, by computing the MLEs on the training data and computing the Bayes factor from the validation data, our models come from *outside* the evaluation data, hence we have a legitimate Bayes factor.

The Bayes factor (5.1) is based on a single random split, which means our conclusions would depend on the specific data split. Using a similar approach to the univariate kernel CVBF method of Hart and Choi (2016), we take N random splits and the resulting $\text{CVBF}_{m,N}$ value is the geometric mean over the N partitions $(\boldsymbol{X}_1^T, \boldsymbol{X}_1^V), \ldots, (\boldsymbol{X}_N^T, \boldsymbol{X}_N^V)$ given by

$$\text{CVBF}_{m,N} = \left[\prod_{k=1}^{N} \text{BF}_m(\mathbf{X}_k^T, \mathbf{X}_k^V) \right]^{1/N}.$$

For the remainder of this chapter, we will refer to $\log \text{CVBF}_{m,N}$ as the CVWE value,

the weight of evidence in favor of the alternative model. As we have seen in the kernel CVBF methods, when we use data splitting to compute the Bayes factor, we pay a penalty in that we must determine what values of m and N to use in practice. We will have recommendations for both of these variables later in this chapter.

5.3 Bayes Factor Consistency Results

Proving consistency of the parametric CVBF method depends on whether or not the two models are nested. Two models are nested when one of the two contains more parameters than the other and when certain parameters of the larger model are set to 0, we obtain the smaller model. For simplicity, we initially consider $\text{BF}_{m,1}$, the Bayes factor for a single random split. We will explore $\text{BF}_{m,N}$ for $N > 1$ in Subsection 5.3.3. We first provide sufficient conditions for consistency when the models are not nested.

5.3.1 Non-Nested Models

Let f_0 continue to be the true underlying density of X_i and assume that there exist parameters $\theta_0 \in \Theta$ and $\lambda_0 \in \Lambda$ such that

$$\int \log f(\mathbf{x}|\theta_0) f_0(\mathbf{x}) d\mathbf{x} = \sup_{\theta \in \Theta} \int \log f(\mathbf{x}|\theta) f_0(\mathbf{x}) d\mathbf{x}$$

and

$$\int \log g(\mathbf{x}|\lambda_0) f_0(\mathbf{x}) d\mathbf{x} = \sup_{\lambda \in \Lambda} \int \log g(\mathbf{x}|\lambda) f_0(\mathbf{x}) d\mathbf{x}.$$

In the non-nested case, we assume that at most one of the two models is correct, meaning that the model contains f_0. However, there still exists the possibility that neither model is correct. When both models are incorrect, we would like the Bayes factor to favor the model that is closer to the truth in a Kullback-Leiber sense. Without loss of generality,

assume that

$$D = \int \log\Big(\frac{f_0(\mathbf{x})}{g(\mathbf{x}|\lambda_0)}\Big) f_0(\mathbf{x}) d\mathbf{x} - \int \log\Big(\frac{f_0(\mathbf{x})}{f(\mathbf{x}|\theta_0)}\Big) f_0(\mathbf{x}) d\mathbf{x} < 0, \tag{5.2}$$

or that the Kullback-Leibler divergence between f_0 and $g(\cdot|\lambda_0)$ is less than that between f_0 and $f(\cdot|\theta_0)$. The conditions for Bayes consistency in the non-nested case are established in the following theorem.

Theorem 5.1. *Assume that the following conditions hold:*

A1. On the basis of a random sample from f_0, the MLEs of θ and λ converge in probability to θ_0 and λ_0, respectively, as sample size tends to ∞.

A2. As a function of the parameter, each of $\log f(\mathbf{x}|\theta)$ and $\log g(\mathbf{x}|\lambda)$ satisfies a Hölder condition for each $\mathbf{x}$. Specifically, there exist functions A and B and positive numbers α_1 and α_2 such that

$$|\log f(\mathbf{x}|\theta_1) - \log f(\mathbf{x}|\theta_2)| \leq A(\mathbf{x})||\theta_1 - \theta_2||^{\alpha_1}$$

for all $\mathbf{x}$ and all $\theta_1, \theta_2 \in \Theta$, and

$$|\log g(\mathbf{x}|\lambda_1) - \log g(\mathbf{x}|\lambda_2)| \leq B(\mathbf{x})||\lambda_1 - \lambda_2||^{\alpha_2}$$

for all $\mathbf{x}$ and all $\lambda_1, \lambda_2 \in \Lambda$.

A3. The integrals $\int A(\mathbf{x}) f_0(\mathbf{x})\, d\mathbf{x}$ and $\int B(\mathbf{x}) f_0(\mathbf{x})\, d\mathbf{x}$ exist finite.

A4. The training set size m tends to ∞ and is bounded by pn for some $p \in (0, 1)$.

If, in addition to A1-A4, (5.2) holds, then

$$\log \boldsymbol{BF}_{m,1}(\boldsymbol{X}^T, \boldsymbol{X}^V) = (n-m)[D + o_p(1)]$$

as $n \to \infty$.

To give a simple proof of Theorem 5.1, we can write the CVBF value based on a single random split as

$$\begin{aligned}\log(\mathrm{BF}_{m,1}(\boldsymbol{X}^T, \boldsymbol{X}^V)) &= \sum_{j=m+1}^{n} \log f(X_j|\hat{\theta}_m) - \sum_{j=m+1}^{n} \log g(X_j|\hat{\lambda}_m) \\ &= (n-m)[D + \delta_1 + \delta_2 + \delta_3 + \delta_4],\end{aligned}$$

where

$$\delta_1 = \frac{1}{n-m}\sum_{j=m+1}^{n} \log f(X_j|\theta_0) - \int \log f(\mathbf{x}|\theta_0) f_0(\mathbf{x}) d\mathbf{x},$$

$$\delta_2 = \frac{1}{n-m}\sum_{j=m+1}^{n} \log f(X_j|\hat{\theta}_m) - \frac{1}{n-m}\sum_{j=m+1}^{n} \log f(X_j|\theta_0),$$

$$\delta_3 = \int \log g(\mathbf{x}|\lambda_0) f_0(\mathbf{x}) d\mathbf{x} - \frac{1}{n-m}\sum_{j=m+1}^{n} \log g(X_j|\lambda_0),$$

and

$$\delta_4 = \frac{1}{n-m}\sum_{j=m+1}^{n} \log g(X_j|\lambda_0) - \frac{1}{n-m}\sum_{j=m+1}^{n} \log g(X_j|\hat{\lambda}_m).$$

We simply need to show that $\delta_i = o_p(1)$ for $i = 1, 2, 3, 4$. By the weak law of large numbers, δ_1 and δ_3 are both $o_p(1)$ since $n - m \to \infty$. Then, by assumption, $\delta_2 = o_p(1)$

since

$$
\begin{aligned}
|\delta_2| &\leq \frac{1}{n-m} \sum_{j=m+1}^{n} |\log f(X_j|\hat{\theta}_m) - \log f(X_j|\theta_0)| \\
&\leq \frac{1}{n-m} \sum_{j=m+1}^{n} A(X_j)||\hat{\theta}_m - \theta_0||^{\alpha_1},
\end{aligned}
$$

which converges to 0 in probability since $\hat{\theta}_m$ is consistent for θ_0 and $(n-m)^{-1} \sum_{j=m+1}^{n} A(X_j)$ converges to its finite expectation. Using a similar argument, $\delta_4 = o_p(1)$ and we reach the desired result.

Therefore, the Bayes factor is bounded by $\exp(n(1-p)[D+o_p(1)])$, which implies that it converges to 0 at an exponential rate. Notice that (5.2) assumed the null model M_1 was true, or at least is closer to f_0 in a Kullback-Leibler sense. We just as easily could assume M_2 to be the true model, in which the result of Theorem 5.1 would still hold. However, since $D > 0$ under the alternative, the Bayes factor converges to ∞ at an exponential rate. Thus, consistency holds under both the null and alternative model. It is also interesting that the Bayes factor is consistent for a training set size that is a fixed proportion of n. In contrast, we saw in Chapter 4 that the kernel CVBF method requires $m = o(n)$. We will see the same requirement for m in the next subsection when we examine consistency for nested models.

5.3.2 Nested Models

In the case of nested models, suppose that $q < p$ and define $r = p - q$. Assume that Λ is the set of all q-vectors $(\theta_1, \ldots, \theta_q)$ such that $(\theta_1, \ldots, \theta_q, 0, \ldots, 0) \in \Theta$. Also assume that M_1 is a subset of M_2 in the sense that $g(\cdot|\lambda) \equiv f(\cdot|(\lambda, 0 \ldots, 0))$ for each $\lambda \in \Lambda$.

Before we provide the conditions for consistency when the smaller model (M_1) is true, we first define some notation. Let k denote the size of a random sample $X_1, \ldots, X_k$. Define $\ell_k(\theta) = \sum_{i=1}^{k} \log f(X_i|\theta), \theta \in \Theta$ to be the corresponding log-likelihood. Assuming

the existence of derivatives, define $\dot{\ell}_k(\theta)$ to be the p-dimensional score vector with ith element

$$\frac{\partial \ell_k(\theta)}{\partial \theta_i}, \quad i = 1, \ldots, p,$$

and $\ddot{\ell}_k(\theta)$ to be the $p \times p$ Hessian matrix having (i, j) element

$$\frac{\partial^2 \ell_k(\theta)}{\partial \theta_i \theta_j}, \quad i = 1, \ldots, p, \ j = 1, \ldots, p.$$

Theorem 5.2 contains the consistency results when the smaller of two nested models is true.

Theorem 5.2. *Assume that the following conditions hold:*

A5. The true density is $f(\cdot|\theta_0)$, where θ_0 is an interior point of Θ and of the form $\theta_0 = (\lambda, 0, \ldots, 0)$ for some $\lambda \in \Lambda$.

A6. The likelihood ℓ_k admits the following Taylor series expansion:

$$\ell_k(\theta) = \ell_k(\theta_0) + (\theta - \theta_0)^T \dot{\ell}_k(\theta) + \frac{1}{2}(\theta - \theta_0)^T \ddot{\ell}_k(\tilde{\theta})(\theta - \theta_0),$$

where $\theta \in \Theta$ and $||\tilde{\theta} - \theta_0|| \leq ||\theta - \theta_0||$.

A7. Let $\hat{\theta}_k$ be the maximizer of $\ell_k(\theta)$ and $\hat{\lambda}$ the q-vector that maximizes $\ell_k(\lambda, 0, \ldots, 0)$ with respect to λ. Then $\hat{\theta}_k$ and $\hat{\theta}_{k,0} = (\hat{\lambda}, 0, \ldots, 0)$ are $\sqrt{k}$-consistent for θ_0 as $k \to \infty$.

A8. For any sequence $\hat{\theta}$ that converges in probability to θ_0, $-\ddot{\ell}_k(\hat{\theta})/k$ is consistent for the Fisher information matrix $I(\theta_0)$ as $k \to \infty$.

If m tends to ∞ with $m = o(n)$, then

$$\log BF_{m,1}(\boldsymbol{X}^T, \boldsymbol{X}^V) = -\frac{n}{2m}\chi^2_{m,r} + o_p\left(\frac{n}{m}\right),$$

where $\chi^2_{m,r}$ converges in distribution to a random variable having the chi-squared distribution with r degrees of freedom.

Proof of Theorem 5.2. Let the parameter vector be $\theta^T = (\theta_1, \ldots, \theta_p)$ and let $\dot{\ell}_{n-m}(\theta)$ be the $p \times 1$ score vector for the validation data and $\ddot{\ell}_{n-m}(\theta)$ the $p \times p$ Hessian matrix for the validation data. Using the Taylor series expansions in (*A6*), we can write $\log\left(\text{BF}_{m,1}(\boldsymbol{X}^T, \boldsymbol{X}^V)\right) = \ell_{n-m}(\theta) - \ell_{n-m}(\lambda)$ as

$$\begin{aligned}\log\left(\text{BF}_{m,1}(\boldsymbol{X}^T, \boldsymbol{X}^V)\right) &= (\hat{\theta}_m - \theta_0)^T \dot{\ell}_{n-m}(\theta_0) + \frac{1}{2}(\hat{\theta}_m - \theta_0)^T \ddot{\ell}_{n-m}(\theta_{1m})(\hat{\theta}_m - \theta_0) - \\ &\quad (\hat{\theta}_{m,0} - \theta_0)^T \dot{\ell}_{n-m}(\theta_0) - \frac{1}{2}(\hat{\theta}_{m,0} - \theta_0)^T \ddot{\ell}_{n-m}(\theta_{2m})(\hat{\theta}_{m,0} - \theta_0),\end{aligned}$$

where $||\theta_{1m} - \theta_0|| \leq ||\hat{\theta}_m - \theta_0||$ and $||\theta_{2m} - \theta_0|| \leq ||\hat{\theta}_{m,0} - \theta_0||$.

Since $\text{E}(\dot{\ell}_{n-m}(\theta_0)) = 0$ and $m = o(n)$, $\dot{\ell}_{n-m}(\theta_0)$ is $\sqrt{n}$-consistent. Combining this fact with Assumption (*A7*), both $(\hat{\theta}_m - \theta_0)^T \dot{\ell}_{n-m}(\theta_0)$ and $(\hat{\theta}_{m,0} - \theta_0)^T \dot{\ell}_{n-m}(\theta_0)$ are $O_p(\sqrt{n/m})$. Therefore, using Assumptions (*A7*) and (*A8*),

$$\begin{aligned}\log\left(\text{BF}_{m,1}(\boldsymbol{X}^T, \boldsymbol{X}^V)\right) &= -\frac{n}{2}\Big[(\hat{\theta}_m - \theta_0)^T I(\theta_0)(\hat{\theta}_m - \theta_0) - \\ &\quad (\hat{\theta}_{m,0} - \theta_0)^T I(\theta_0)(\hat{\theta}_{m,0} - \theta_0)\Big] + o_p\left(\frac{n}{m}\right).\end{aligned} \tag{5.3}$$

Consider partitioning $I(\theta_0)$ as follows: $I(\theta_0) = \begin{bmatrix} I_{11} & I_{12} \\ I_{12}^T & I_{22} \end{bmatrix}$, where I_{11} is $q \times q$, I_{12} is $q \times r$ and I_{22} is $r \times r$. Now define $A_{p\times p} = \begin{bmatrix} I_{11}^{-1} & 0_{q\times r} \\ 0_{r\times q} & 0_{r\times r} \end{bmatrix}$. From p. 231 of van der Vaart (1998), we have the following two equations.

$$\sqrt{m}(\hat{\theta}_m - \theta_0) = \frac{1}{\sqrt{m}} I(\theta_0)^{-1} \dot{\ell}_m(\theta_0) + o_p(1)$$

and

$$\sqrt{m}(\hat{\theta}_{m,0} - \theta_0) = \frac{1}{\sqrt{m}} A \dot{\ell}_m(\theta_0) + o_p(1),$$

where $\dot{\ell}_m(\theta)$ is the score vector for the training data. Substitution of the last two expressions into (5.3) yields

$$\log\left(\text{BF}_{m,1}(\boldsymbol{X}^T, \boldsymbol{X}^V)\right) = -\left(\frac{n}{2m}\right)\frac{1}{\sqrt{m}}\dot{\ell}_m(\theta_0)^T\left[I(\theta_0)^{-1} - A\right]\frac{1}{\sqrt{m}}\dot{\ell}_m(\theta_0) + o_p\left(\frac{n}{m}\right).$$

We will utilize Result 5.15 on p. 112 in Monahan (2008), to show that

$$\frac{1}{\sqrt{m}}\dot{\ell}_m(\theta_0)^T\left[I(\theta_0)^{-1} - A\right]\frac{1}{\sqrt{m}}\dot{\ell}_m(\theta_0) \xrightarrow{D} \chi^2_r$$

by verifying that $B = [I(\theta_0)^{-1} - A]\, I(\theta_0)$ is idempotent and of rank r, since by the multivariate Central Limit Theorem, $m^{-1/2}\dot{\ell}_m(\theta_0) \xrightarrow{D} N_p(0, I(\theta_0))$.

Let $\mathbf{I}_k$ denote the $k \times k$ identity matrix. Then,

$$\begin{aligned}
B^2 &= \left[I(\theta_0)^{-1} - A\right] I(\theta_0) \left[I(\theta_0)^{-1} - A\right] I(\theta_0) \\
&= \left[\mathbf{I}_p - AI(\theta_0)\right] \left[I(\theta_0)^{-1} - A\right] I(\theta_0) \\
&= \left[I(\theta_0)^{-1} - 2A + AI(\theta_0)A\right] I(\theta_0) \\
&= \left[I(\theta_0)^{-1} - A\right] I(\theta_0),
\end{aligned}$$

and thus B is idempotent. Finally, the rank of B can be determined by examining the rows of

$$\begin{aligned}
\left[I(\theta_0)^{-1} - A\right] I(\theta_0) &= \mathbf{I}_p - \begin{bmatrix} I_{11}^{-1} & 0_{q\times r} \\ 0_{r\times q} & 0_{r\times r} \end{bmatrix} \begin{bmatrix} I_{11} & I_{12} \\ I_{12}^T & I_{22} \end{bmatrix} \\
&= \mathbf{I}_p - \begin{bmatrix} \mathbf{I}_q & I_{11}^{-1}I_{12} \\ 0_{r\times q} & 0_{r\times r} \end{bmatrix} \\
&= \begin{bmatrix} 0_{q\times q} & -I_{11}^{-1}I_{12} \\ 0_{r\times q} & \mathbf{I}_r \end{bmatrix}.
\end{aligned}$$

The last r rows are certainly linearly independent, but the first q rows are linear combina-

tions of the last r rows. Hence, the rank of B is indeed r and the proof is complete.

Theorem 5.2 shows that under standard regularity conditions found in likelihood theory, the CVBF value for a single random split is Bayes consistent at an exponential rate under the null hypothesis. As seen in Johnson and Rossell (2010), under these same conditions, a standard Bayes factor would converge to 0 at the rate $n^{-r/2}$ when the smaller model (null hypothesis) is true. However, provided that m increases with n sufficiently slowly, the CVBF will be bounded in probability by $\exp(-n^{\alpha})$ for α arbitrarily close to 1.

Notice how the Bayes factor depends on a chi-square random variable with r degrees of freedom, asymptotically. If we were to test these nested hypotheses from a frequentist perspective, when the null hypothesis is true, $2[\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_{n,0})] \rightarrow \chi^2_r$ according to Wilks (1938) where $\hat{\theta}_{n,0}$ and $\hat{\theta}_n$ are the constrained and unconstrained MLEs from the entire data set. Therefore, both the parametric CVBF and the standard likelihood ratio test for the same nested hypotheses depend on the same chi-squared random variable when the smaller model is true. The difference of course is the $-n/m$ term in the Bayes factor. It is because of this factor that we have the convergence to 0. Since the likelihood ratio statistic is always at least 1, it is not an odds ratio and will not be consistent if we used it as a Bayes factor. Remarkably, it is due to the data splitting and formulation of the two simple models from the training data that lead to the consistency of the likelihood ratio under the null hypothesis.

Lastly, the following theorem gives the conditions for consistency when the larger model (alternative hypothesis) is true. We do not prove this result as it follows similar arguments as in the proof of Theorem 5.1.

Theorem 5.3. *Assume that the model is identifiable in the sense that* $D_{KL}(\theta_1, \theta_2) > 0$ *for all* $\theta_1 \neq \theta_2 \in \Theta$*, where*

$$D_{KL}(\theta_1, \theta_2) = \int f(\mathbf{x}|\theta_1) \log \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_2)} d\mathbf{x}.$$

Let the true parameter value be θ_1, which is such that at least one of its last r components is nonzero. Let λ_0 (which is assumed to exist) be the element of Λ that maximizes $\int f(\mathbf{x}|\theta_1) \log f(\mathbf{x}|(\lambda, 0, \ldots, 0))d\mathbf{x}$ with respect to λ. If

$$\frac{1}{n-m} \sum_{j=m+1}^{n} \log f(X_j|\hat{\theta}_m) \quad \textit{and} \quad \frac{1}{n-m} \sum_{j=m+1}^{n} \log f(X_j|\hat{\theta}_{m,0})$$

are consistent for

$$\int f(\mathbf{x}|\theta_1) \log f(\mathbf{x}|\theta_1)d\mathbf{x} \quad \textit{and} \quad \int f(\mathbf{x}|\theta_1) \log f(\mathbf{x}|(\lambda_0, 0, \ldots, 0))d\mathbf{x},$$

respectively, then as $n - m \to \infty$

$$\log \boldsymbol{BF}_{m,1}(\boldsymbol{X}^T, \boldsymbol{X}^V) = (n-m)D_{KL}(\theta_1, (\lambda_0, 0, \ldots, 0)) + o_p(n-m).$$

According to Theorem 5.3, the Bayes factor is asymptotic (in probability) to $\exp(Cn)$ for some positive constant C and thus converges to ∞ as $n \to \infty$ when the alternative hypothesis is true. As in the case where the smaller model is true, it is necessary for both m and $n - m$ to tend to ∞ when the larger model is true. However, we may allow m to be a fixed fraction of n such that $m = pn$ for $0 < p < 1$. Combining these two rules for m, we can use the following rule for choosing m in nested models: "Let m be the largest integer smaller than $n/2$ that produces desirable behavior of the Bayes factor when the smaller model is true." The term "desirable behavior" means that the Bayes factor is less than $1/20$ with probability close to 1 under the null hypothesis. This rule will help us determine m using either of the calibration methods described in previous chapters.

5.3.3 The Benefit of Multiple Data Splits

As we have seen repeatedly thus far, we typically choose between $30 \leq N \leq 50$ random splits of the data when computing the overall CVBF value. Most of the justification has been based on empirical evidence, but in the case of nested models, we can directly see the effect theoretically. For a single random split, expression (5.3) shows that when the smaller model is true, the dominant (random) term $\chi^2_{m,r}$ depends completely on the training data. Let $N = n/m$, which is chosen to be an integer for convenience, and consider the N data splits for which the training sets are $\boldsymbol{X}_i^T = (X_{(i-1)m+1}, \ldots, X_{im}), i = 1, \ldots, N$. According to Theorem 5.2, the log-Bayes factor for the ith of these splits has the form

$$\log \text{BF}_{m,1}(\boldsymbol{X}_i^T, \boldsymbol{X}_i^V) = -\frac{n}{2m}\chi^2_{m,r,i} + o_p\Big(\frac{n}{m}\Big),$$

where $\chi^2_{m,r,i}$ depends only on $\boldsymbol{X}_i^T$. Since, the $N \to \infty$ training sets are independent of each other, it follows that

$$\frac{1}{N}\sum_{i=1}^{N} \log \text{BF}_{m,1}(\boldsymbol{X}_i^T, \boldsymbol{X}_i^V) = -\frac{n}{2m}r + o_p\Big(\frac{n}{m}\Big).$$

Thus, the random noise due to the χ^2 random variable can be completely removed by averaging over independent splits. Also, this gives us a more definitive approach for how to choose the form and number of the N data splits in the case of nested hypotheses. For a given sample size n, take $N = n/m$ and then the training sets are N independent partitions of the observed sample. In small sample cases this may not be the wisest approach since using this scheme can result in a very small number of training sets. For instance if $n = 500$ and $m = 100$, then the resulting CVBF value is based on only $N = 5$ random splits.

In order to compare the performance of the parametric CVBF method when using the independent training set approach to the more typical choice of $N \geq 30$ dependent random

splits, we consider testing a normal model against a skew-normal alternative for univariate data. Under the null model, we sample data from the standard normal distribution $N(0,1)$ and under the alternative, the data come from a $SN(0,1,10)$ model. The training set sizes are taken to be $m = 100, 125, 200, 250,$ and 400 for $n = 500, 1000, 2000, 5000,$ and 10000, respectively. These training set sizes are chosen such that $m/n \to 0$ and $m \to \infty$ which are conditions in Theorems 2 and 3. Also, n/m is an integer for our convenience.

For each data set in the case of dependent training sets we compute the CVWE value based on $N = 5, 8, 10, 15, 20, 25, 40, 60, 80,$ and 100 random splits. We draw 500 independent random samples from each of the two models. The relative effect of using independent and dependent training sets is seen in Table 5.1, where for each n, n/m and N are the same.

Truth	Type of Split	$n = 500$ $m = 100$	$n = 1000$ $m = 125$	$n = 2000$ $m = 200$	$n = 5000$ $m = 250$	$n = 10000$ $m = 400$
$N(0,1)$	I	$-2.3(3.0)$	$-4.0(3.9)$	$-4.6(3.6)$	$-10.1(5.2)$	$-11.9(5.4)$
	D	$-2.3(3.4)$	$-4.1(3.9)$	$-5.0(3.8)$	$-10.0(5.2)$	$-12.0(5.5)$
$SN(0,1,10)$	I	$22.5(168.9)$	$71.9(167.0)$	$205.0(46.5)$	$557.5(60.2)$	$1170.2(55.5)$
	D	$29.8(169.7)$	$75.1(159.5)$	$207.2(39.3)$	$559.1(60.3)$	$1171.0(50.9)$

Table 5.1: Median CVWE values (with interquartile ranges) for 500 replications of testing normal against skew-normal densities. The CVWE values are obtained from $N = n/m$ independent (I) or dependent (D) training sets.

Notice in Table 5.1 that it really does not make a difference whether we compute the CVWE value on independent or dependent training sets. Also, under the normal (null) model, we expect the the CVWE value to be approximately $-\frac{n}{2m}$ since $r = 1$. This is

what we see from this small simulation for both types of splits. For ease of construction and computation, we will simply take $N \geq 30$ dependent random splits since the results are essentially the same.

5.4 Simulation Studies

In this section, we carry out a series of simulation studies to explore the performance of the parametric CVBF method and compare it to standard frequentist and Bayesian methods. In Subsection 5.4.1 we test the fit of a univariate exponential model against gamma model alternatives and examine the choice of training set size. Next, we look at comparing the normal and skew-normal models for trivariate data in Subsection 5.4.2, which turns out to be a more difficult problem than might be expected. We also compare the parametric CVBF method to the standard frequentist t-test (Subsection 5.4.3) and to a traditional Bayes factor approach (Subsection 5.4.4) in a simple linear regression context.

5.4.1 Testing the Fit of a Univariate Exponential Versus Gamma Model

To investigate the effect of m, we will test an exponential density against a gamma alternative. Letting gamma(α, β) denote a gamma density with shape parameter α and rate parameter β, data were generated from three densities: gamma($1/2, 2$), gamma($1, 2$) (exponential), and gamma($2, 2$). Three sample sizes, $n = 100, 500,$ and 1000 were considered for each gamma density and CVWE values were computed for m in $\{.05n, .10n, \ldots, .5n\}$ and $N = 50$. The simulation results are provided in Figure 5.1.

In order to make the scale of the plots more informative, we have used the transformation $t(\text{CVWE}) = \text{sgn}(\text{CVWE})|\text{CVWE}|^{1/2}$, where sgn($u$) is the sign of u. The dashed horizontal line at $\pm\sqrt{\log 20}$ in each plot represents the strong evidence threshold from Kass and Raftery (1995). When the null model is true (top panel of Figure 5.1) for the gamma(1,2) distribution, the CVWE values decrease monotonically as $m \to 0$. This is what we have seen under the null in every CVWE (either kernel or parametric) scenario.

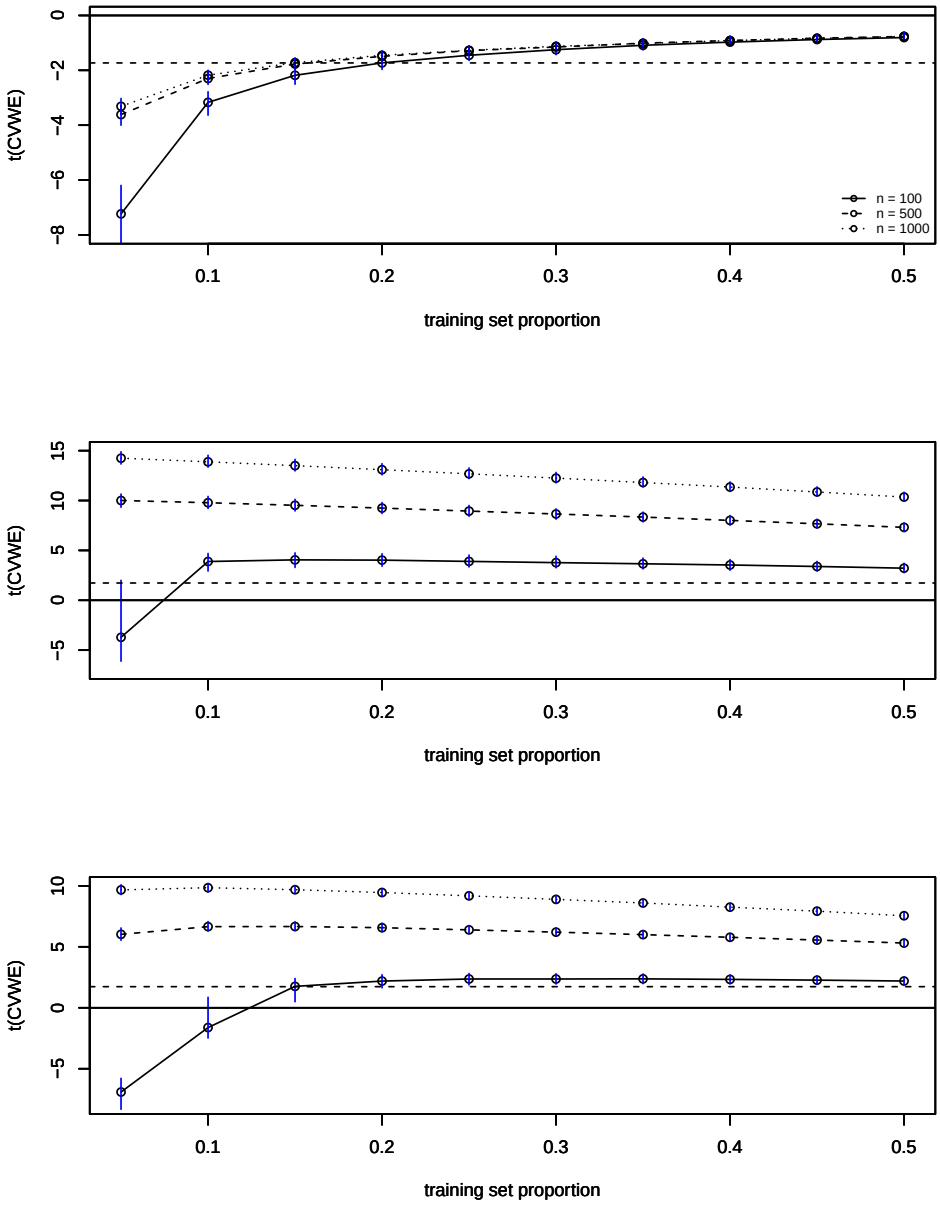

Figure 5.1: Median of transformed CVWE when testing exponential versus gamma densities. Results are based on 1,000 replications from gamma(1,2) (*top panel*), gamma(1/2,2) (*middle panel*), and gamma(2,2) (*bottom panel*) densities. The solid, dashed and dotted lines correspond to $n = 100, 500,$ and 1000, respectively. The upper and lower ends of the vertical lines indicate quartiles, and the dashed horizontal line indicates strong evidence according to the scale of Kass and Raftery (1995).

Any training set size $m \leq 0.15n$ for sample size n would produce CVWE values that indicate strong evidence in favor of the exponential model. Though not included in the plots, all training set sizes provide positive evidence in favor of the exponential model.

Under the alternative models (gamma$(1/2, 2)$ and gamma$(2, 2)$), provided that $n \geq 500$, any training set size between $.05n$ and $.5n$ will produce (with very high probability) a CVWE value that indicates strong evidence against the exponential model. For smaller sample sizes, like $n = 100$ here, the choice of m becomes more important. For instance, in Figure 5.1 when $\alpha = 1/2$ (middle panel) we need to choose $m \geq 0.1n$ and when $\alpha = 2$ (bottom panel) the training set size needs to be even larger with $m \geq 0.2n$. Based on our intuition and the results of this simulation, m_n/n must be larger for smaller n where m_n is the ideal choice of m under the alternative model for given n. Overall, since the parametric CVBF performs adequately under the null, we can use our advice from Subsection 5.3.2 and take m to be a larger proportion of n, especially when n is small.

5.4.2 Testing Trivariate Normality Versus Skew-Normality

The parametric CVBF method is well-suited for applications to multivariate data provided that we can compute the necessary MLEs. Here we consider the example of testing normality against a skew-normal alternative for trivariate data. On the surface it seems that this situation could be easily handled using Bayesian methods, but in fact, it is rather difficult. In the typical (ξ, Ω, α) parameterization, a singularity exists in the Hessian matrix when the skew parameter α is a 0-vector. Therefore, to perform a standard Bayesian hypothesis test, one could reparameterize the skew-normal model and follow the population Monte Carlo approach using objective priors, as developed by Liseo and Parisi (2013). Unfortunately, this approach becomes very complicated beyond two dimensions. In contrast, the parametric CVBF method easily handles this hypothesis test in all dimensions.

For sample sizes $n = 1000, 2500,$ and 5000, we draw 256 independent sample from the

trivariate standard normal distribution and the trivariate skew-normal distribution with parameters $\xi = \mathbf{0}$, $\Omega = \mathbf{I}_3$, and $\alpha = \mathbf{10}$. For each data set we use $m = 0.1n, 0.2n, 0.3n, 0.4n,$ and $0.5n$ and $N = 50$. The results are summarized in Figure 5.2 using the same transformation of the CVWE values that was utilized in Subsection 5.4.1.

When the data are sampled from the trivariate standard normal distribution, the parametric CVBF method finds strong evidence in favor of the normal model when $m \leq 0.3n$ and positive evidence for all training set sizes. For sample sizes $n \geq 2500$, the CVWE values for skew-normal data indicate overwhelming evidence in favor of the skew-normal model for any training set size. When $n = 1,000$, the training set needs to contain at least $m = 200$ observations before we find positive evidence in favor of the skew-normal model. As the dimension increases, we simply need more observations to adequately estimate the (quadratically) increasing number of parameters. However, provided that we have a large enough sample, these results extend for dimension greater than 3. Thus, the parametric CVBF approach can make quick and easy work of a difficult hypothesis test.

How does the parametric CVBF method compare to the scaled $\text{CVBF}_K(\mathcal{S})$ method for testing trivariate normality for normal and skew-normal data. Back in Subsection 4.4.4, we conducted a similar test in three dimensions using data from non-standard distributions. Specifically, the normal distribution had parameters: $\mu = (3.4, 5.5, 3.5)^T$ and $\mathbf{\Sigma} = \begin{bmatrix} 5.5 & 2.1 & -.2 \\ 2.1 & 2.0 & .02 \\ -.2 & .02 & 9.9 \end{bmatrix}$ and the skew normal distribution had parameters: $\xi = (-14.1, 18.9, 15.5)^T$, $\mathbf{\Omega} = \begin{bmatrix} 5.5 & -3.9 & 1.3 \\ -3.9 & 5.1 & -1.6 \\ 1.3 & -1.6 & 2.1 \end{bmatrix}$, and $\alpha = (15.9, 7.1, -6.0)^T$.

For 100 random samples with $n = 1,000$ observations from both of these distributions, we computed the parametric and kernel CVBF methods using $N = 28$ random splits and training set sizes $m = 100, 200, 300, 400,$ and 500. The resulting median CVWE values over the 100 samples are provided in Table 5.2.

The results from this simulation are very interesting. For normal data, both the parametric and kernel CVBF methods find at least positive evidence in favor of the normal

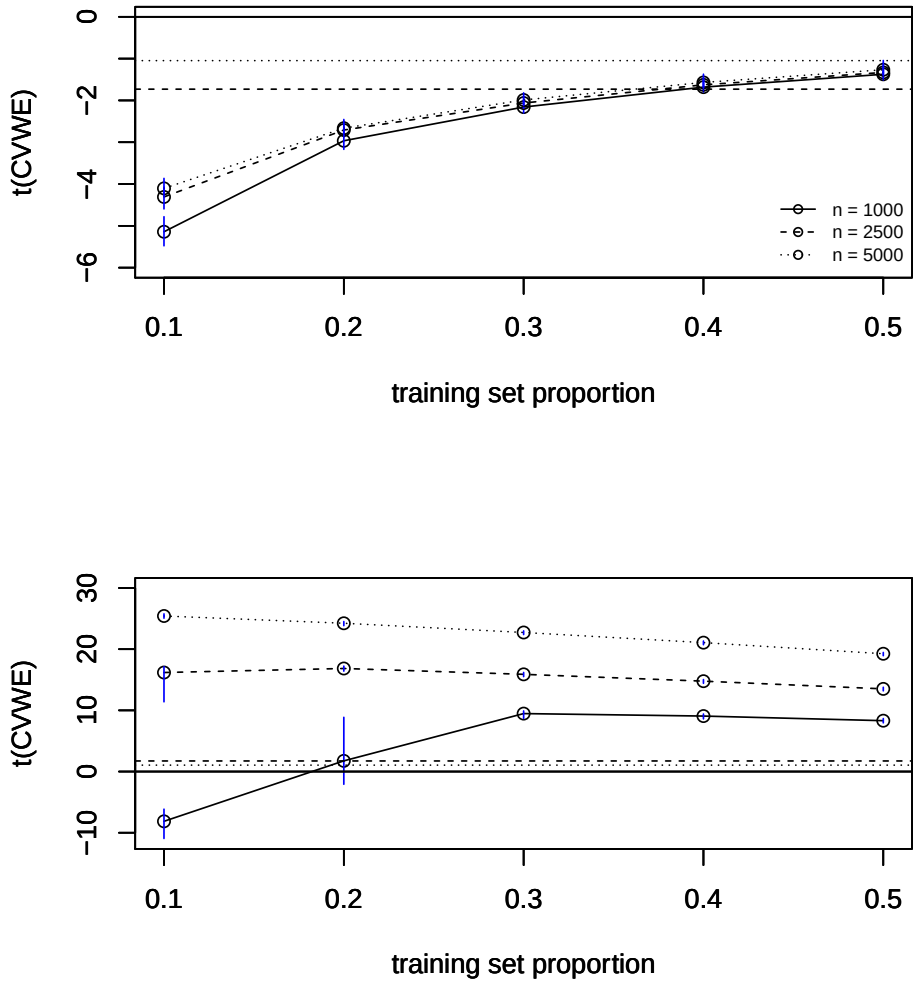

Figure 5.2: Median of transformed CVWE when testing trivariate normality for 256 samples from $N(\mathbf{0}, \mathbf{I}_3)$ (*top panel*) and $SN(\mathbf{0}, \mathbf{I}_3, \mathbf{10})$ (*bottom panel*) data. The solid, dashed and dotted lines correspond to $n = 1000, 2500$ and 5000, respectively. The upper and lower ends of the vertical lines indicate quartiles, and the dashed and dotted horizontal lines indicate strong and positive evidence according to the scale of Kass and Raftery (1995).

Method	Model	$m = 100$	$m = 200$	$m = 300$	$m = 400$	$m = 500$
CVBF_P	Normal	-26.6	-9.0	-4.5	-2.9	-1.7
	Skew-Normal	-37.5	67.7	79.3	74.6	62.5
CVBF_K	Normal	-84.1	-46.2	-28.5	-16.9	-7.6
	Skew-Normal	-34.9	6.5	22.5	29.9	31.9

Table 5.2: Median CVWE_P and scaled $\text{CVWE}_K(\mathcal{S})$ values for 100 random samples of size $n = 1,000$ from either a trivariate normal or skew-normal model using training set sizes $m = 100, 200, 300, 400,$ and 500 and $N = 28$ random splits.

model. However, the kernel CVWE values indicate far stronger (overwhelming) evidence across all training set sizes. When the alternative model is true, the two methods reverse roles in that the parametric CVBF method finds magnitudes more evidence against the normal model compared to the kernel approach.

These results are exactly what we expect to see. In the parametric CVBF method, we are comparing two nested parametric models with only $n = 1,000$ observations using a likelihood ratio of simple models estimated from the training data. In the kernel CVBF method, we are using a Bayes factor to compare a parametric model to a nonparametric model, where the marginal likelihoods serve as model averages over their respective parameters. Therefore, when the null model is true, it is not surprising to see the larger CVBF values for the parametric approach compared to the kernel approach. The parametric model in the kernel approach will look markedly better compared to the nonparametric model, whereas both the estimated skew-normal and normal models will be harder to distinguish. When the skew normal model is true, the parametric approach should produce larger CVBF values since the estimated skew-normal model will fit the data far better than the estimated normal model. In the kernel approach, at least one member of the alternative model should be closer to the skew-normal model, but we know from a Kullback-Leibler

sense that the kernel model is often closer to a normal model. Therefore, the skew-normal model should be more difficult to distinguish.

5.4.3 Comparing CVBF$_P$ to a Frequentist Test

When comparing a frequentist test and a Bayesian test for the same hypothesis testing problem, we have described how the significance level of $\alpha = .05$ is often too liberal and should tend to 0 in order to agree with the Bayesian test. In fact, in the simulations of Section 4.7, we showed that the frequentist tests for multivariate normality all had Type I error rates near $\alpha = .05$, whereas for the same data sets, the Type I error rate for the kernel CVBF method was 0 for appropriate training set size.

To explore this scenario once again, consider testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ in a simple linear regression setting. The model we consider is such that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent with $Y_i|(X_i = x) \sim N(\beta_0 + \beta_1 x, \sigma^2)$ and $X_i \sim N(\mu, \sigma_X^2),\ i = 1, \ldots, n$. Take $X_1, \ldots, X_n, Y_1, \ldots, Y_n \overset{iid}{\sim} N(0, 1)$, in which the null hypothesis is true, and sample 10,000 data sets of size $n = 1,000$.

For each data set we compute the P-value from the classical t-test on $\hat{\beta}_1$, the least squares estimate of the slope parameter. More specifically, the P-value is equal to

$$2P(|\hat{t}| > t_{n-2}) \quad \text{where} \quad \hat{t} = \frac{\hat{\beta}_1}{(\hat{\sigma}^2 / \sum_{i=1}^n X_i^2)},$$

with

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(X_i - \bar{X}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}, \quad \hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2, \quad \text{and}$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n.$$

The parameter estimates and sample means are computed on all n pairs of observations

for the t-test. As for the CVBF method, we only use the training data, $m = 100$ in this simulation, to compute the estimates and means. For a single random split, the Bayes factor we compute is given by

$$\text{BF}_{m,1} = \frac{\hat{\sigma}_1^{-(n-m)} \exp\Big(-\frac{1}{2\hat{\sigma}_1^2} \sum_{j=m+1}^{n} (Y_j - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \Big)}{\hat{\sigma}_0^{-(n-m)} \exp\Big(-\frac{1}{2\hat{\sigma}_0^2} \sum_{j=m+1}^{n} (Y_j - \bar{Y}_m)^2 \Big)}$$

where $\hat{\sigma}_0^2 = m^{-1} \sum_{i=1}^{m} (Y_i - \bar{Y}_m)^2$. The resulting CVWE value is the average of values of the form $\log(\text{BF}_{m,1})$ over 50 random splits.

Over the 10,000 data sets, 94.86% of the CVWE values were less than $-\log 20$, which indicates strong evidence in favor of the correct null model. As for the t-test, as expected roughly 5% (5.04% to be exact) of the data sets produce P-values less than $\alpha = .05$. In order to see that a level 0.05 test is too liberal, in 268 of the 504 t-tests that produced Type I errors, the corresponding CVWE value finds strong evidence in favor of the null model. That means that in over 53% of data sets where the frequentist makes a Type I error, we can actually find strong evidence for the null model. Similarly, in 466 (92.5%) of the 504 data sets, we would find positive evidence in favor of the null model using the parametric CVBF method. This situation will only continue to be more disturbing as the sample size increases because of the consistency results of Theorem 5.2. In fact, $P(\text{CVWE} < -\log(20) | P \leq \alpha) \to 1$ as $n \to \infty$ for any fixed α.

The CVWE and P-value tend to agree when the CVWE values are very large. Suppose we rejected the null model when the CVWE value was greater than $\log 3$, which occurred only 3 times in the 10,000 data sets. This is fairly liberal when it comes to odds ratios because a log-odds ratio must be greater than 0 for the alternative model to be favored. For the t-test to have the same Type I error rate α would be 0.0003. In fact, using this significance level, 4 data sets produce significant P-values, and in 2 of these data sets

the Bayesian would also reject the null. Since $P(\text{CVWE} > \log(3)) \to 0$ as $n \to \infty$, a necessary condition for the frequentist and Bayes tests to agree closely in terms of Type I errors is that $\alpha \to 0$ as $n \to \infty$.

5.4.4 Comparing CVBF$_P$ to a Traditional Bayes Factor

In a traditional Bayesian hypothesis test where the two models are nested, the Bayes factor is typically consistent at a slower than exponential rate when the smaller model is true. Theorem 5.2 proves that the parametric CVBF method is consistent at an exponential rate when the null hypothesis is true. To explore this large sample property, we compare the parametric CVWE values to traditional log Bayes factors in a simple regression setting. Consider testing the following hypotheses:

$$H_0 : Y_i = \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$
$$H_1 : Y_i = X_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad X_i \sim N(0, 1), \quad \beta \neq 0,$$

where X_i and ϵ_i are independent. We assume that the null model is true, i.e., that $\beta = 0$.

The parametric CVBF approach closely follows the computations in the previous subsection, with the extra caveat that we do not consider the intercept parameter β_0. Thus the CVWE value for a single random split of the observed data pairs (X_i, Y_i) and training sample size m is given by

$$\log(\text{BF}_{m,1}) = \frac{n-m}{2} \log\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right) - \frac{\sum_{j=m+1}^{n}(Y_j - \hat{\beta}X_j)^2}{2\hat{\sigma}_1^2} + \frac{\sum_{j=m+1}^{n} Y_j^2}{2\hat{\sigma}_0^2},$$

where $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$, and $\hat{\beta}$ are computed from the training data as follows:

$$\hat{\beta} = \frac{\sum_{i=1}^{m} X_i Y_i}{\sum_{i=1}^{m} X_i^2}, \quad \hat{\sigma}_0^2 = \frac{1}{m}\sum_{i=1}^{m} Y_i^2, \quad \text{and} \quad \hat{\sigma}_1^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{\beta}X_i)^2.$$

The overall CVWE value is the arithmetic average of values of the form $\log(\text{BF}_{m,1})$ across $N = 50$ splits.

As for the traditional Bayes factor, we require prior distributions for the parameters under both the null and alternative hypotheses. The only unknown parameter in the null model is σ^2, so we take $\sigma^2 \sim$ inverse-gamma$\left(\frac{1}{2}, (2n)^{-1}\sum_{i=1}^{n} Y_i^2\right)$, which is a conjugate reference prior. For the alternative model, Hoff (2009) provides UIR priors for $\beta|\sigma^2$ and σ^2: $\beta \sim N\left(\hat{\beta}, n\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$ and $\sigma^2 \sim$ inverse-gamma$\Big(\frac{1}{2}, (2n)^{-1}[\boldsymbol{Y} - \hat{\beta}\boldsymbol{X}]^T[\boldsymbol{Y} - \hat{\beta}\boldsymbol{X}]\Big)$. Use of these priors leads to marginal likelihoods that can be computed analytically resulting in the following Bayes factor

$$\text{BF} = \frac{\Big([\boldsymbol{Y} - \hat{\beta}\boldsymbol{X}]^T[\boldsymbol{Y} - \hat{\beta}\boldsymbol{X}]\Big)^{1/2}\Big[\boldsymbol{Y}^T\boldsymbol{Y}\left(\frac{2n+1}{2n}\right) + \hat{\beta}^2\boldsymbol{X}^T\boldsymbol{X}\left(\frac{3}{2n}\right) - \hat{\beta}\boldsymbol{Y}^T\boldsymbol{X}\Big]^{-\frac{n+1}{2}}}{\Big(\boldsymbol{Y}^T\boldsymbol{Y}\Big)^{1/2}\Big[\boldsymbol{Y}^T\boldsymbol{Y}\left(\frac{2n+1}{2n}\right)\Big]^{-(n+1)/2}}.$$

To compare the traditional Bayes factor to the parametric CVBF method, we simulate 10,000 data sets such that $X_1, \ldots, X_n, Y_1, \ldots, Y_n \overset{iid}{\sim} N(0,1)$. The sample sizes we consider are $n = 1000, 5000, 10000, 25000, 50000,$ and 100000 with respective training set sizes of $m = 250, 500, 750, 1000, 1500,$ and 2000. The pairs (m, n) are chosen such that both $m, n \to \infty$ and $m/n \to 0$ as required in Theorem 5.2. For each data set we compute the log-Bayes factor and CVWE value and the resulting pairs $(\text{CVWE}, \log(\text{BF}))$ are plotted in Figure 5.3.

The results of Figure 5.3 verify that indeed the traditional log-Bayes factor tends to $-\infty$ at a much slower rate than the parametric CVBF method. Also notice that regardless of sample size, there are data sets where the traditional Bayesian regression approach will incorrectly favor the alternative model. This is unlike the CVBF_P approach where once $n \geq 10,000$, the CVWE value correctly concludes in favor of the null model in *all* 10,000 data sets.

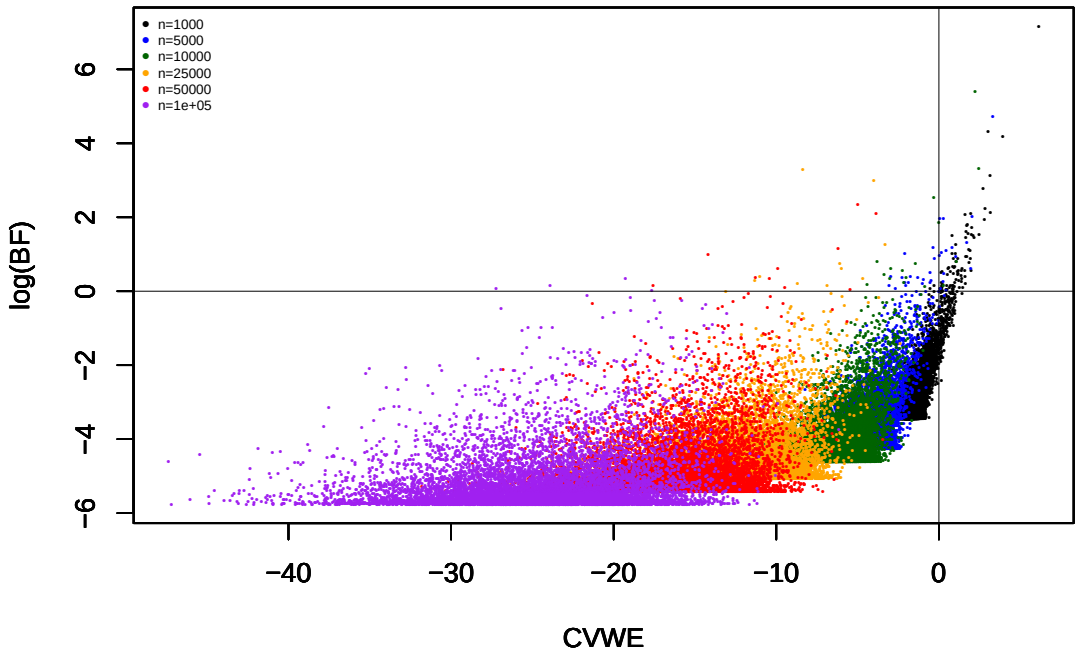

Figure 5.3: Comparison of the parametric CVWE values and the Bayes factors from a traditional Bayesian regression analysis. Each color represents one of the 6 (n, m) pairs: $(1,000, 250)$, $(5,000, 500)$, $(10,000, 750)$, $(25,000, 1,000)$, $(50,000, 1,500)$, and $(100,000, 2,000)$. Each individual point is one of $10,000$ replications of an (n, m) pair.

5.5 Real Data Analysis

In this section, we apply the parametric CVBF methodology to civil engineering data that are publicly available at the UC Irvine Machine Learning Repository and originally published by Yeh (1998). The data consist of $n = 1030$ determinations of $Y =$ concrete compressive strength under a variety of different settings for the following eight design variables: $x_1 =$ kg cement, $x_2 =$kg blast furnace slag, $x_3 =$ kg fly ash, $x_4 =$ kg water, $x_5 =$ kg superplasticizer, $x_6 =$ kg coarse aggregate, $x_7 =$ kg fine aggregate, and $x_8 =$ age (in days). We consider the following two models, both of which regress Y on all eight design variables. The first model is a Gaussian linear model in which the errors are assumed to be homoscedastic and the second model uses the same linear model, however the errors are heteroscedastic. The errors are assumed to be independent, thus the likelihoods for the two models have standard forms.

The model considered for the mean of Y was linear in $x_1, \ldots, x_8$, and $\sqrt{x_8}$. When the errors are assumed to be homoscedastic, this model has an R^2 value of .820. However, based on the residual plot in Figure 5.4, perhaps this assumption is invalid since the variance of the residuals tends to increase with the mean. The goal is to use the parametric CVBF method to determine if a model allowing for heteroscedastic errors better models these data.

To show that the homoscedastic and heteroscedastic models are indeed nested, consider the following. Let r denote the model for the conditional mean of $Y|\boldsymbol{X} = (x_1, \ldots, x_8)$ given by

$$r(x_1, \ldots, x_8) = \beta_0 + \beta_1 x_1 + \cdots + \beta_8 x_8 + \beta_9 \sqrt{x_8}.$$

If $V(x_1, \ldots, x_8)$ is the variance of an error term when the values of the design variables

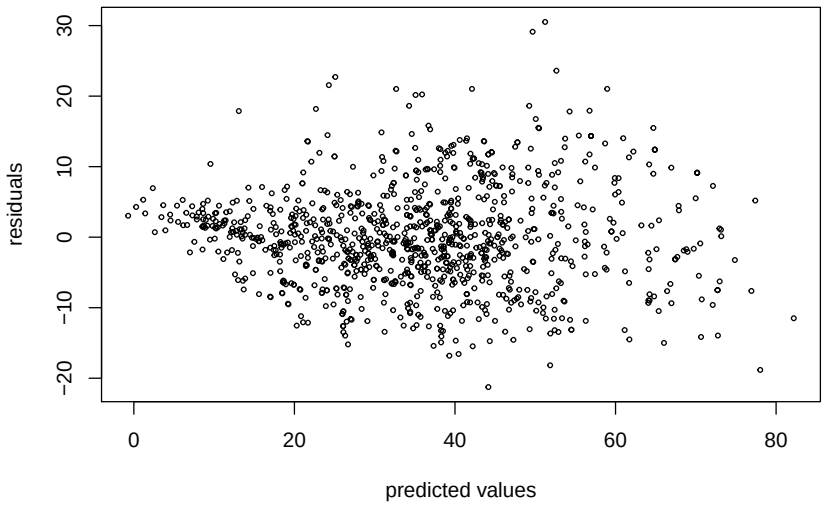

Figure 5.4: Residuals from homoscedastic linear model fitted to the civil engineering data.

are $x_1, \ldots, x_8$, then we take

$$V(x_1, \ldots, x_8) = \exp(a_0 + a_1 r(x_1, \ldots, x_8)),$$

where a_0 and a_1 are unknown parameters and when $a_1 = 0$ we obtain the homoscedastic model.

Under the null model, the MLEs for the error variance and slope parameters are easily obtained for the homoscedastic model using ordinary least squares regression. As for the heteroscedastic model, the parameters a_0, a_1, and β are determined through maximization of the log-likelihood function for the weighted least squares regression model. Using training sample sizes $m = 100, 200, 300, 400,$ and 500 and $N = 200$ random splits we compute the CVWE value from the civil engineering data. The resulting median and quartiles of the 200 CVWE values at each training set size are in the top panel of Figure

5.5. Certainly there appears to be strong evidence in favor of the heteroscedastic model. While the interquartile range is very large (and extends below 0) for $m = 100$, any other choice of m results in overwhelming evidence for non-constant variance. To choose m, we employ a calibration technique where we randomly sample 1,000 data sets (each of size 1030) from the fitted homoscedastic linear model. For each of the 1,000 data sets, we compute the CVWE values at the same five training set sizes and $N = 50$ random splits. The resulting medians and quartiles from the null data are provided in the bottom panel of Figure 5.5. Based on our recommendations for choosing m when the smaller of two nested models is true, $m = 200$ is a suitable choice here. If the homoscedastic model were indeed the true model, then it would be extremely unlikely to see the median CVWE value that we observed when $m = 200$.

5.6 Summary and Conclusions

In order to compare two parametric models with a Bayes factor, using cross-validation Bayes factors proves to be a very simple and intuitive approach that also has excellent large sample properties. The methodology in Section 5.2 is straightforward, but contains some subtle yet important details. First, by selecting the models from outside the data upon which we evaluate the Bayes factor, the likelihood ratio *is* a valid Bayes factor. Without the data splitting, the classical likelihood ratio would be inconsistent as a Bayes factor. Also, while we can use a likelihood ratio test from the frequentist perspective for non-nested models, there are a few philosophical problems that arise. For instance, we may need to arbitrarily choose which of our two models is the null model and the entire formulation of the test will depend on this choice. Also, the nice result of Wilks (1938) will no longer apply and thus the asymptotic distribution of the likelihood ratio will no longer be χ^2. However, since the likelihood ratio in the parametric CVBF methodology truly is a Bayes factor, we can apply Bayesian hypothesis testing methods which easily handle

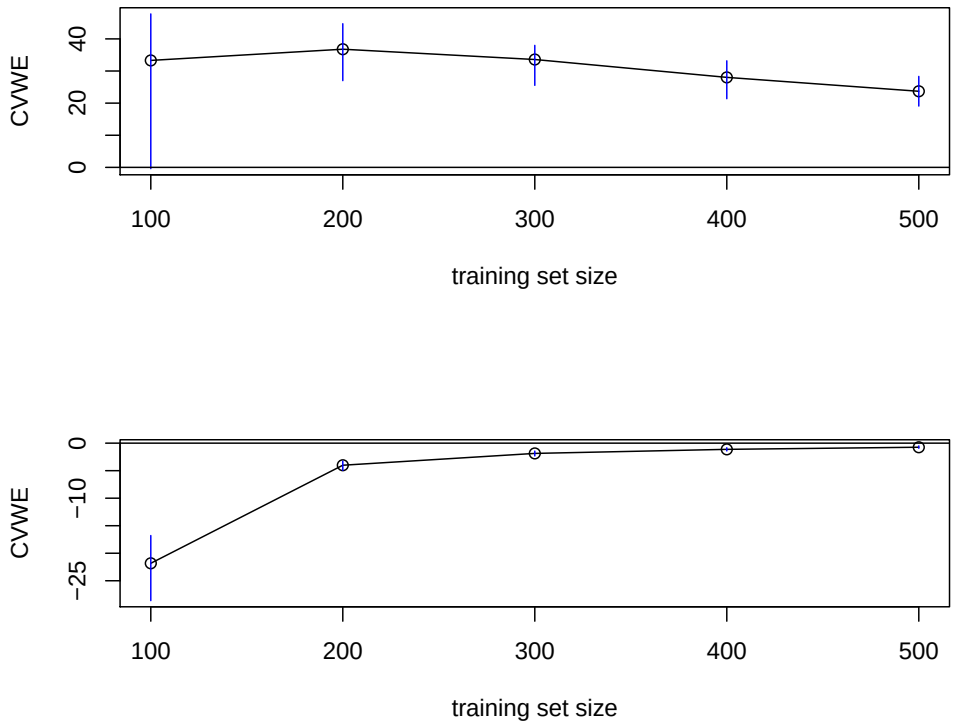

Figure 5.5: At each training set size, the median and quartiles for the CVWE values from the observed civil engineering data based on 200 random splits (*top panel*) and for the 1,000 data sets from the estimated homoscedastic model with 50 splits (*bottom panel*) are provided.

these difficulties. Lastly, instead of formulating proper prior distributions and numerically integrating the marginal likelihoods, we only need to estimate unknown parameters from the training data and evaluate the likelihood function on the validation data.

In Section 5.3, we provided the conditions required for the parametric CVBF method to be consistent at an exponential rate for both nested and non-nested models, regardless of which model is correct. This is superior to typical Bayes factors which often converge at a slower rate when the smaller of two nested models is true. In order to use the data splitting technique, we must determine m and N. However, the conditions in Theorems 5.1-5.3 give us some guidance for choosing the number and form of the random splits N and how to select m using calibration.

The simulations in Section 5.4 and the data analysis in Section 5.5 really illustrate the superiority of the parametric CVBF method to its frequentist and Bayesian counterparts. First, the parametric CVBF method makes the difficult Bayesian problem of testing multivariate normality versus skew-normality extremely easy in any dimension. Next, in the simple linear regression setting, the significance level of the classical t-test for the slope parameter must tend to 0 as $n \to \infty$ for it to agree with the parametric CVBF method. Also, we see the far superior convergence and Type I error rates of the parametric CVBF compared to the traditional Bayes factor approach. Finally, in the real data analysis we see how useful the parametric CVBF method is in testing for heteroscedasticity in a linear regression model. There are frequentist tests for this same problem, i.e. Breusch and Pagan (1979), but they suffer from the same significance level and Type I error rate problems of frequentist testing as $n \to \infty$.

6. SUMMARY AND FUTURE WORK

The next natural course of action is to combine the kernel and parametric CVBF methods into a new hybrid approach for goodness-of-fit testing. In the kernel CVBF method, we use the training data to fit the kernel model and then compute the marginal likelihood by integrating over the smoothing parameter space. Suppose that instead of going through the formal Bayesian approach to find the marginal likelihood, we determine the optimal smoothing parameter for the kernel density estimate on the training data and consider that our simple model to use as the alternative model. In essence, we use the methodology from the parametric CVBF method and instead of having two parametric models, we have a parametric null and a nonparametric alternative. On the training data, we determine the nonparametric model that best fits the data as well as the MLEs for the parametric model. Those become the two simple models from which we can compute the likelihood ratio on the validation data.

By considering this hybrid approach, we may be able to combine the computational simplicity of the parametric CVBF approach with more sophisticated nonparametric density estimation techniques. Take for instance the full bandwidth matrix version of the multivariate kernel density estimate. We saw just how complicated the evaluation of the likelihood can be even in only two dimensions. Now, for testing multivariate normality, the hybrid methodology would be as follows. First, form the training and validation data sets $\boldsymbol{X}^T$ and $\boldsymbol{X}^V$, respectively by randomly splitting the data. Under the parametric normal model $f_d(\cdot|\mu, \mathbf{\Sigma})$, compute $\hat{\mu} = m^{-1}\sum_{i=1}^{m} X_i$ and $\hat{\mathbf{\Sigma}} = m^{-1}\sum_{i=1}^{m}[X_i - \hat{\mu}][X_i - \hat{\mu}]^T$ from the training data. As for the nonparametric kernel model $\hat{f}_d(\cdot|\mathbf{H})$, determine the optimal bandwidth matrix $\mathbf{H}_{opt}$ using Zhang et al. (2006) on the training data. Now, we have our

two simple models for which we can compute the Bayes factor (likelihood ratio) given by

$$\text{BF}_{m,1} = \frac{\prod_{j=m+1}^{n} \hat{f}_d(X_j|\mathbf{H}_{opt}, \boldsymbol{X}^T)}{\prod_{j=m+1}^{n} f_d(X_j|\hat{\mu}, \hat{\boldsymbol{\Sigma}})}.$$

Now, we avoid the problem of maximization and numerical integration over the constrained space of symmetric positive definite matrices. Just in this simple example, the hybrid approach has great potential in terms of computational simplicity.

Another advantage to using the hybrid CVBF is we can adopt more sophisticated density estimation techniques that are both faster and provide more accurate representations of the true density compared to the simple kernel density estimate. In Chapter 4, we only considered the multivariate kernel density estimate with a single smoothing matrix to make finding the prior distribution easy and minimize the number of unknown parameters. However, provided that the training data has a sufficient number of observations to fit the nonparametric model of our choosing, we can essentially use *any* density estimation technique as our alternative model. One possible simple extension would be to consider adaptive or variable bandwidth kernel density estimates (Breiman et al. (1977), Silverman (1986), and Terrell and Scott (1992)). This would be a first step toward better estimation of the density in regions with few observations. For further improvement on the kernel density estimate we could consider the *fastKDE* method of O'Brien et al. (2016) which represents the kernel density estimate as the product of the empirical characteristic function and the inverse Fourier transform of the kernel. We could also take a method from machine learning called *BoostKDE* (Di Marzio and Taylor, 2005) which is another iterative procedure that begins with the multivariate kernel density estimate as the initial estimate, updates the weights for each data vector (originally $w_i = 1/n$), and then multiplies (and renormalizes) the M estimates. Certainly, these extensions of the basic kernel density estimate would be far too complicated to use in the kernel CVBF method for a

variety of reasons, all of which no longer exist in the hybrid approach.

We can also consider density estimation techniques that are not kernel methods (see Izenman (1991) for a brief overview). In the univariate case, we could use any orthogonal series or basis expansion method such as wavelets, B-splines, Fourier series, or polynomials as our nonparametric estimate. For a cursory look at all of these expansions see Ramsay and Silverman (2005). For any dimension we could let the the alternative model be a finite mixture of normal distributions. The mixing proportion, mean vector, and covariance matrix for each of the component normal distributions can be determined using the Expectation-Maximization algorithm on the training data. Another possible method is projection pursuit density estimation described by Friedman et al. (1984) which is an iterative updating procedure beginning with a proposed parametric density and multiplying it by a series of univariate kernel estimates. A very interesting approach specific to finding a way around the curse of dimensionality was proposed by Nagler and Czado (2016) who examine simplified vine copulas noting that the joint d-dimensional density can be written as a decomposition into marginal densities and bivariate copula densities. The authors argue that under certain conditions, their density estimate achieves a rate of convergence equal to the rate of a two-dimensional estimator regardless of d, ergo the curse of dimensionality ceases to exist.

One final method that we considered in the kernel CVBF method for an improvement over the typical kernel density estimate is a semiparametric one first created by Hjort and Jones (1996) and then extended to multivariate data by Jarnicka (2009). The Hjort-Jones (HJ) estimator is very simple intuitively and is written as,

$$\hat{f}(\mathbf{x}) = \frac{f_{\text{init}}(\mathbf{x})\tilde{f}(\mathbf{x})}{(K_d * f_{\text{init}})(\mathbf{x})},$$

where $\tilde{f}(\cdot)$ is the multivariate kernel density estimate, $f_{\text{init}}(\cdot)$ is a parametric distribution,

and $*$ denotes convolution. This estimate can be thought of as a nonparametric start with a parametric correction (or vice versa). Regardless if the parametric model f_{init} fits the data well, the resulting estimate will still be reasonable. This is due to the property that as $h \to \infty$, $\hat{f} \to f_{\text{init}}$ and as $h \to 0$, $\hat{f} \to \tilde{f}$. Therefore, if f_{init} is completely wrong, the selected bandwidth will be small, resulting in a near fully nonparametric estimate. If the parametric model is correct, the bandwidth will be extremely large and the final estimate will be fully parametric.

One very concerning detail in using the HJ estimator in this hybrid approach occurs when the null model is true. By selecting the null model to be f_{init}, under the null, the likelihood ratio may be 1 since the MLEs are computed from the same training data. Certainly, this would prove to be worthless as a Bayes factor, but perhaps certain modifications could be used. For instance, let f_{init} be a different parametric model that may be plausible. This would be the ultimate hybrid CVBF approach since we have two competing parametric models that we can simultaneously test along with a nonparametric model.

Of course this list of techniques is by no means an exhaustive list of methods for multivariate density estimation that we could apply to this new hybrid CVBF approach. However, it does show the vast number of research avenues that we can take in the future. Certainly some of these methods may not prove worthwhile due to complexity, consistency, convergence rates, formulation, etc. But we can imagine that one of the more sophisticated methods for multivariate density estimations will help alleviate the curse of dimensionality allowing us to apply the CVBF method to higher dimensional data beyond $d = 10$. The combination of the kernel and parametric CVBF methods appears to be a very fruitful area of further research.

REFERENCES

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society Series B (Methodological) 53*(1), 111–142.

Andrews, D. and A. Herzberg (1985). *Data: a collection of problems from many fields for the student and research worker*. Springer Series in Statistics. New York: Springer-Verlag.

Azzalini, A. and A. Capitanio (1998, February). Statistical applications of the multivariate skew-normal distribution. arXiv:0911.2093.

Basu, S. and S. Chib (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association 98*, 224–235.

Berger, J. and M. Delampady (1987). Testing precise hypotheses. *Statistical Science 2*(3), 317–335.

Berger, J. and A. Guglielmi (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association 96*(453), 174–184.

Berger, J. and L. Perrichi (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association 91*(433), 109–122.

Berger, J. and T. Sellke (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association 82*(397), 112–122.

Bhattacharya, A. and J. D. Hart (2016, August). Partitioned cross-validation for divide-and-conquer density estimation. arXiv:1609.00065.

Biau, G. and A. Mas (2010, March). PCA-kernel estimation. arXiv:1003.5089.

Breiman, L., W. Meisel, and E. Purcell (1977). Variable kernel estimates of multivariate densities. *Technometrics 19*(2), 135–144.

Breusch, T. and A. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica 47*(5), 1287–1294.

Carota, C. and A. Parmigiani (1996). On Bayes factors for nonparametric alternatives. *Bayesian statistics 5*, 507–511.

Chang, X., S. Lin, and Y. Wang (2016, March). Divide and conquer local average regression. arXiv:1601.06239.

Chib, S. and T. Kuffner (2016, July). Bayes factor consistency. arXiv:1607.00292.

Chiu, S. and K. Liu (2009). Generalized Cramér-von Mises goodness-of-fit tests for multivariate distributions. *Computational Statistics and Data Analysis 53*(11), 3817–3834.

Conigliani, C., J. Castro, and A. O'Hagan (2000). Bayesian assessment of goodness of fit against nonparametric alternatives. *The Canadian Journal of Statistics 28*(2), 327–342.

D'Agostino, R. and M. Stephens (1986). *Goodness-of-fit techniques*, Volume 68 of *Statistics: textbooks and monographs*. New York: Marcel Dekker.

Davidson, L. A., D. V. Nguyen, R. M. Hokanson, E. S. Callaway, R. B. Isett, N. D. Turner, E. R. Dougherty, N. Wang, J. R. Lupton, R. J. Carroll, and R. S. Chapkin (2004). Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. *Cancer Research 64*, 6797–6804.

Davis, P. and P. Rabinowitz (2007). *Methods of Numerical Integration* (Second ed.). Mineola, N.Y.: Dover Publications.

de Bruijn, N. (1961). *Asymptotic methods in analysis* (Second ed.). Bibliotheca mathematica, a series of monographs on pure and applied mathematics: v. 4. New York: Interscience Publishers.

Delampady, M. and J. Berger (1990). Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit. *The Annals of Statistics 18*(3), 1295–1316.

Di Marzio, M. and C. Taylor (2005). On boosting kernel density methods for multivariate

data: density estimation and classification. *Statistical Methods & Applications 14*, 163–178.

Duong, T. and M. Hazelton (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics 15*(1), 17–30.

Duong, T. and M. Hazelton (2005). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics 32*, 485–506.

Evans, M. and T. Swartz (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science 10*(3), 254–272.

Fodor, I. (2002). A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab.

Friedman, J., W. Stuetzle, and A. Schroeder (1984). Projection pursuit density estimation. *Journal of the American Statistical Association 79*(387), 599–608.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition* (Second ed.). Boston: Academic Press.

Geisser, S. and W. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*(365), 153–160.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. H. Rubin (2014). *Bayesian data analysis* (Third ed.). Texts in Statistical Science. Boca Raton: CRC Press.

Ghosh, J. and R. Ramamoorthi (2003). *Bayesian Nonparametrics*. Springer Series in Statistics. New York: Springer-Verlag.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Annals of Statistics 15*(4), 1491–1519.

Hall, P. and J. Marron (1987). Extent to which least-squares cross-validation minimises integrated squared error in nonparametric density estimation. *Probability Theory and Related Fields 74*, 567–581.

Härdle, W. and D. Scott (1992). Smoothing by weighted average of rounded points. *Com-*

putational Statistics 7, 97–128.

Hart, J. and T. Choi (2016). Nonparametric goodness of fit via cross-validated Bayes factors. *Bayesian Analysis 12*, 653–677.

Hawkins, D. (1981). A new test for multivariate normality and homoscedasticity. *Technometrics 23*(1), 105–110.

Hjort, N. L. and M. Jones (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics 24*(4), 1619–1647.

Hoff, P. (2009). *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. New York: Springer Science + Business Media.

Huber-Carol, C., N. Balakrishnan, M. Nikulin, and M. Mesbah (Eds.) (2002). *Goodness-of-fit tests and model validity*. Statistics for Industry and Technology. New York: Springer Science + Business Media.

Izenman, A. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association 86*(413), 205–224.

Jarnicka, J. (2009). Multivariate kernel density estimation with a parametric support. *Opuscula Mathematica 29*(1), 41–55.

Jeffreys, H. (1961). *Theory of Probability* (Third ed.). Oxford: Clarenden Press.

Johnson, V. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America 110*(48), 19313–19317.

Johnson, V. and D. Rossell (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society B 72*(2), 143–170.

Justel, A., D. Peña, and R. Zamar (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics and Probability Letters 35*(3), 251–259.

Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Korkmaz, S., D. Goksuluk, and G. Zararsiz (2016). MVN: An R package for assessing

multivariate normality. *R Journal 6*(2), 151–162.

Kullback, S. and R. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Lavine, M. and M. Schervish (1999). Bayes factors: What they are and what they are not. *The American Statistician 53*(2), 119–122.

Lehmann, E. and J. Romano (2005). *Testing Statistical Hypotheses* (Third ed.). Springer Texts in Statistics. New York: Springer Science + Business Media.

Li, R., D. K. J. Lin, and B. Li (2013). Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry 29*(5), 399–409.

Lichman, M. (2013). UCI machine learning repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Lindley, D. V. (1957). A statistical paradox. *Biometrika 44*, 187–192.

Liseo, B. and A. Parisi (2013). Bayesian inference for the multivariate skew-normal model: A population Monte Carlo approach. *Computational Statistics and Data Analysis 63*, 125–138.

Lumley, T. (2017). *survey: analysis of complex survey samples.* R package version 3.32.

Malkovich, J. and A. Afifi (1973). On tests for multivariate normality. *Journal of the American Statistical Association 68*(341), 176–179.

Mardia, K. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika 57*(3), 519–530.

McVinish, R., J. Rousseau, and K. Mengersen (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scandinavian Journal of Statistics 36*, 337–354.

Mecklin, C. and D. Mundfrom (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review 72*(1), 123–138.

Monahan, J. (2008). *A Primer on Linear Models*. Texts in Statistical Science. Boca Raton:

CRC Press.

Müller, P. and F. Quintana (2004). Nonparametric bayesian data analysis. *Statistical Science 19*(1), 95–110.

Müller, P., F. Quintana, A. Jara, and T. Hanson (2015). *Bayesian nonparametric data analysis*. Springer Series in Statistics. Switzerland: Springer International Publishing.

Nadarajah, S. and S. Kotz (2005). Matematical properties of the multivariate t distribution. *Acta Applicandae Mathematica 89*(1-3), 53–84.

Nagler, T. and C. Czado (2016, May). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. arXiv:1503.03305.

Neyman, J. (1937). "smooth" tests for goodness of fit. *Scandinavian Actuarial Journal 1937*(3-4), 149–199.

O'Brien, T., K. Kashinath, N. Cavanaugh, and J. Collins, W.D. O'Brien (2016). A fast and objective multidimensional kernel density estimation method: fastKDE. *Computational Statistics and Data Analysis 101*, 148–160.

O'Hagan, A. (1991). Discussion on "posterior bayes factors," by M. Aitkin. *Journal of the Royal Statistical Society Series B (Methodological) 53*(1), 136.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society B (Methodological) 57*(1), 99–138.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5 50*, 157–175.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. New York: Springer Science + Business Media.

Raykar, V., R. Duraiswami, and L. Zhao (2010). Fast computation of kernel estimators.

Journal of Computational and Graphical Statistics 19(1), 205–220.

Rayner, J. and D. Best (1989). *Smooth Tests of Goodness of Fit*. New York: Oxford University Press.

Reiersøl, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica 18*(4), 375–389.

Robert, C. and G. Casella (2004). *Monte Carlo statistical methods* (Second ed.). Springer Texts in Statistics. New York: Springer Science + Business Media.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics 23*(3), 470–473.

Royston, J. (1982). An extension of Shapiro and Wilk's w test for normality in large samples. *Journal of the Royal Statistical Society Series C 31*(2), 115–124.

Ruli, E., N. Sartori, and L. Venture (2016, December). Improved Laplace approximation for marginal likelihoods. arXiv:1502.06440.

Rust, R. and D. Schmittlein (1985). A Bayesian cross-validated likelihood method for comparing alternative specifications of quantitative models. *Markeing Science 4*, 20–40.

Sain, S., K. Baggerly, and D. Scott (1994). Cross-validation of multivariate densities. *Journal of the American Statistical Association 89*(427), 807–817.

Scott, D. (1992). *Multivarate density estimation: theory, practice, and visualization*. Wiley Series in Probability and Mathematical Statistics. New York: New York : John Wiley & Sons.

Scott, D. and J. Thompson (1983). Probability density estimation in higher dimensions. In J. Gentle (Ed.), *Proceedings of the Fifteenth Interface of Computer Science and Statistics*, pp. 173–179.

Scott, D. and M. Wand (1991). Feasibility of multivariate density estimation. *Biometrika 78*(1), 197–205.

Silverman, B. (1986). *Density estimation for statistics and data analysis* (1st ed.). Number 26 in Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Simonoff, J. (1996). *Smoothing methods in statistics*. Springer Series in Statistics. New York: Springer-Verlag.

Tang, Q. and R. Karunamuni (2016). Fast and accurate computation for kernel estimators. *Computational Statistics and Data Analysis 94*, 49–62.

Terrell, G. and D. W. Scott (1992). Variable kernel density estimation. *The Annals of Statistics 20*(3), 1236–1265.

Thode, H. (2002). *Testing for normality*, Volume 164 of *Statistics: textbooks and monographs*. New York: Marcel Dekker.

Tokdar, S. and R. Martin (2013). Bayesian test of normality versus a Dirichlet process mixture alternative. arXiv:1108.2883.

van der Laan, M., S. Dudoit, and S. Keleş (2004). Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology 3*, online publication.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. New York: Cambridge University Press.

Verdinelli, I. and L. Wasserman (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *The Annals of Statistics 26*(4), 1215–1241.

Villasenor Alva, J. and E. González Estrada (2009). A generalization of Shapiro-Wilk's test for multivariate normality. *Communication in Statistics - Theory and Methods 38*(11), 1870–1883.

Wand, M. (1994). Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics 3*(4), 433–445.

Wand, M. and M. Jones (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association 88*(422),

520–528.

Wand, M. and M. Jones (1994). Multivariate plug-in bandwidth selection. *Computational Statistics 9*, 97–116.

Wand, M. and M. Jones (1995). *Kernel Smoothing* (First ed.). Number 60 in Monographs on Statistics and Applied Probability. New York: Chapman & Hall.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Texts in Statistics. New York: Springer.

Wilks, S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics 19*, 60–62.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research 28*(12), 1797–1808.

Zhang, X., M. King, and R. Hyndman (2006). A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics and Data Analysis 50*, 3009–3031.

APPENDIX

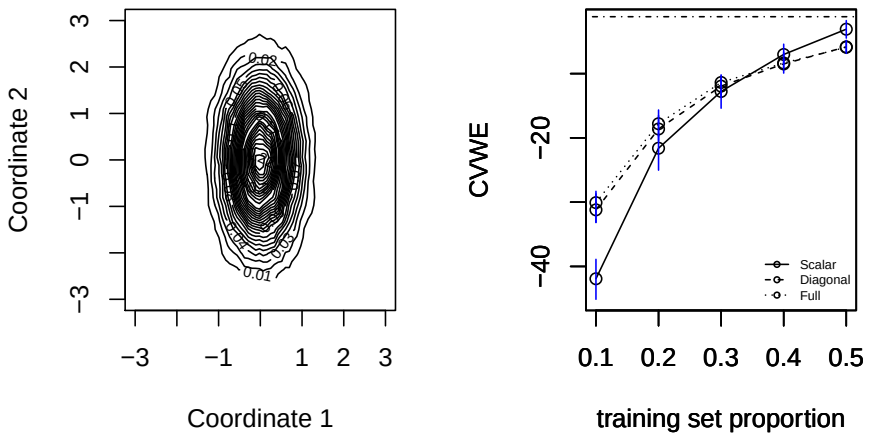

Figure A.1: Testing bivariate normality using $\mathrm{CVWE}_K(\mathcal{S})$, $\mathrm{CVWE}_K(\mathcal{D})$, and $\mathrm{CVWE}_K(\mathcal{F})$ for uncorrelated normal data.

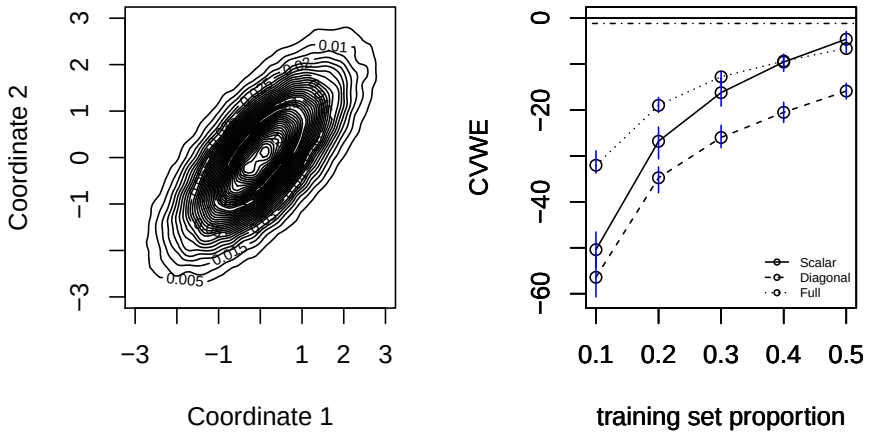

Figure A.2: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for correlated normal data.

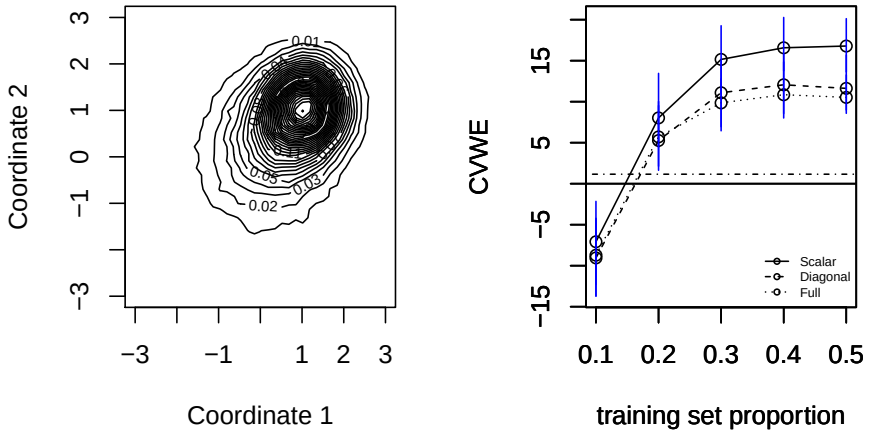

Figure A.3: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for skewed data.

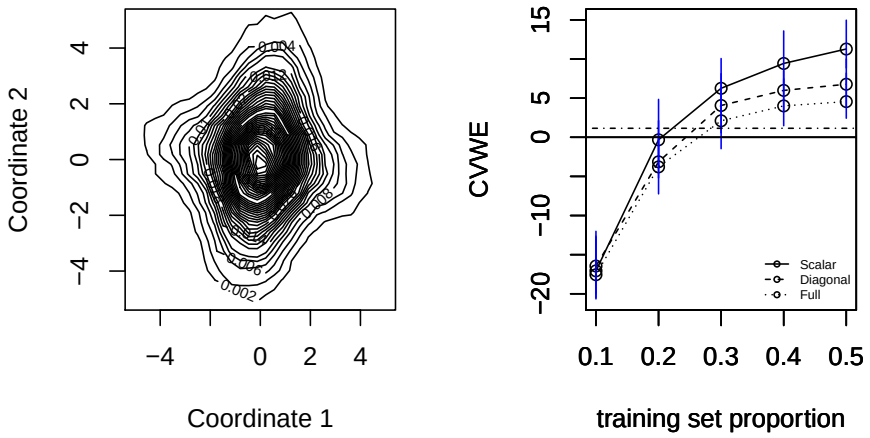

Figure A.4: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for kurtotic data.

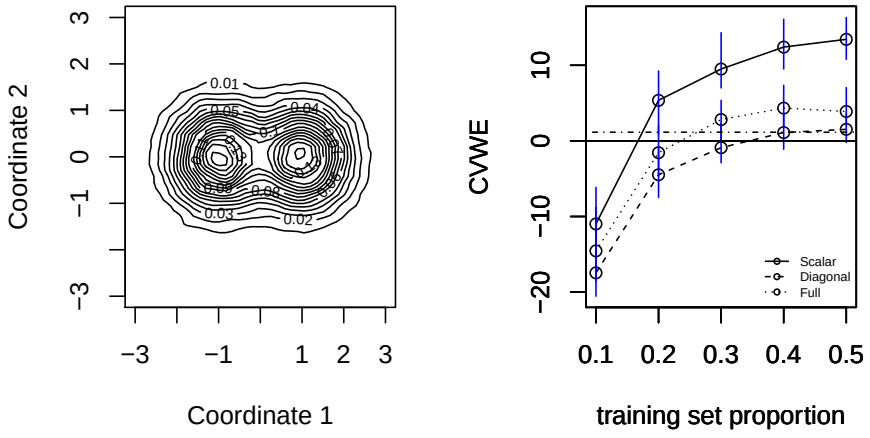

Figure A.5: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for bimodal (I) data.

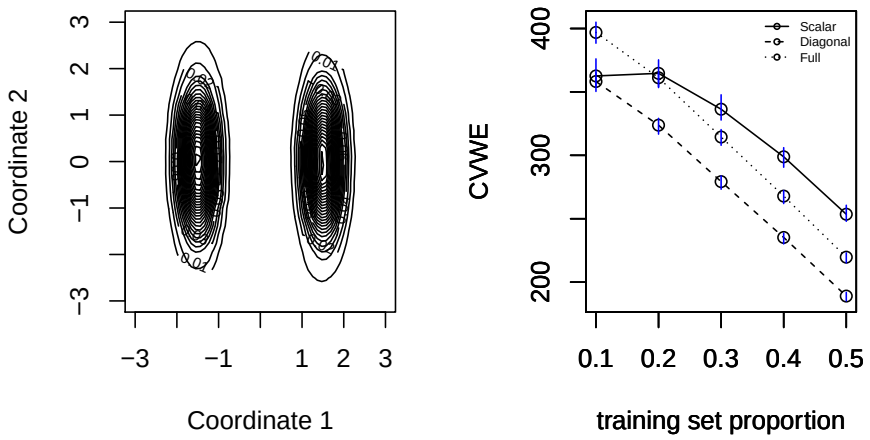

Figure A.6: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for bimodal (II) data.

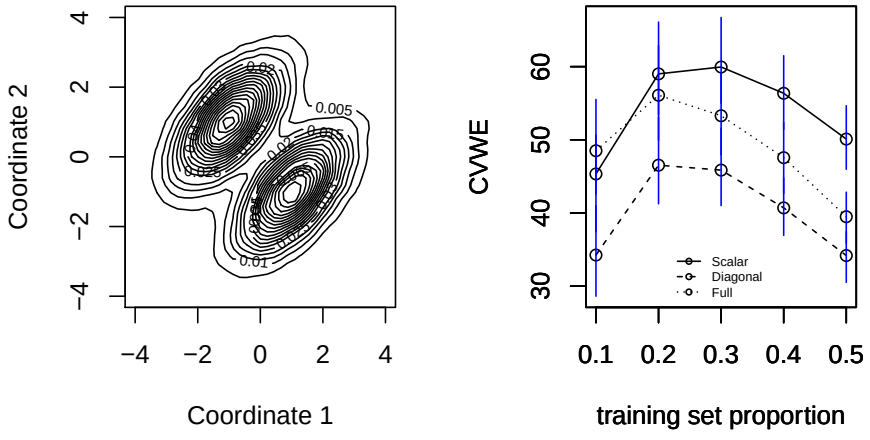

Figure A.7: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for bimodal (III) data.

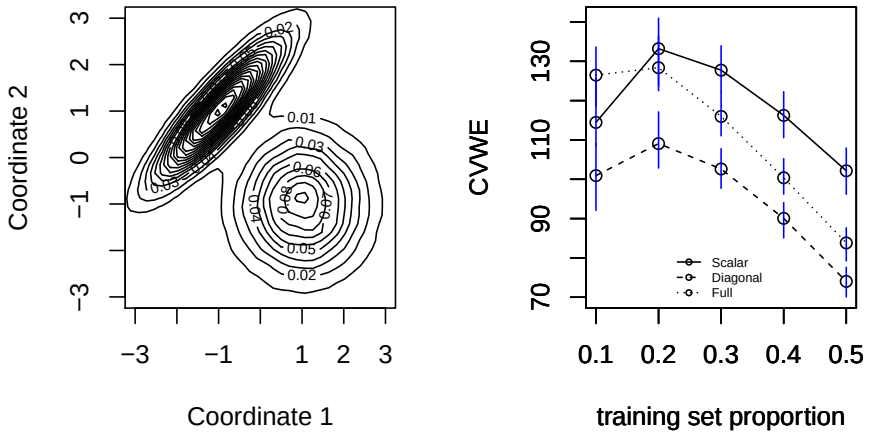

Figure A.8: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for bimodal (IV) data.

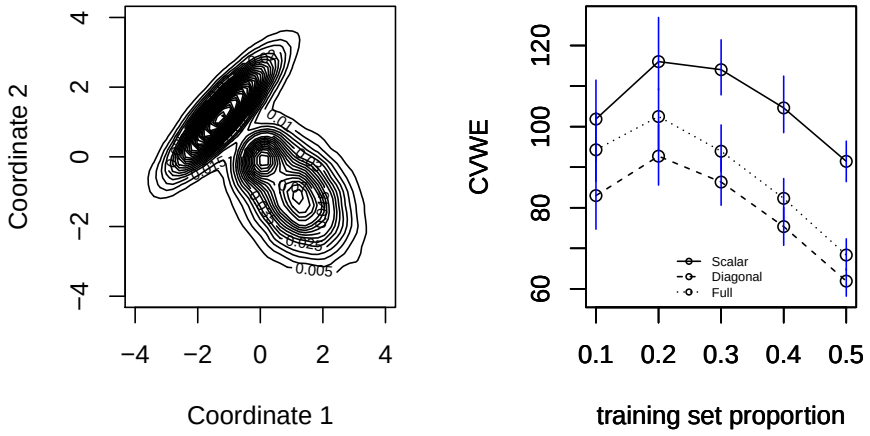

Figure A.9: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for trimodal (I) data.

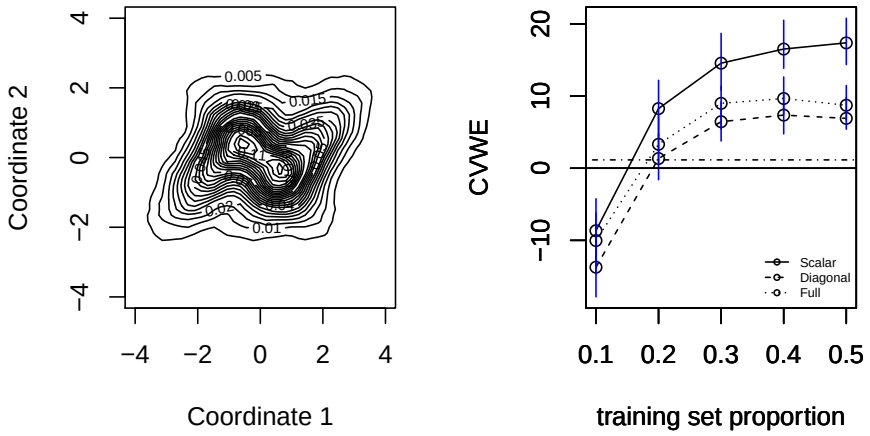

Figure A.10: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for trimodal (II) data.

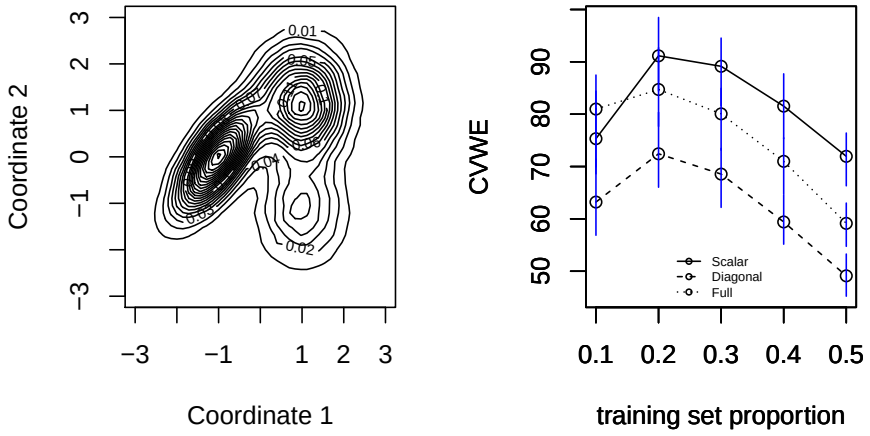

Figure A.11: Testing bivariate normality using $\text{CVWE}_K(\mathcal{S})$, $\text{CVWE}_K(\mathcal{D})$, and $\text{CVWE}_K(\mathcal{F})$ for trimodal (III) data.

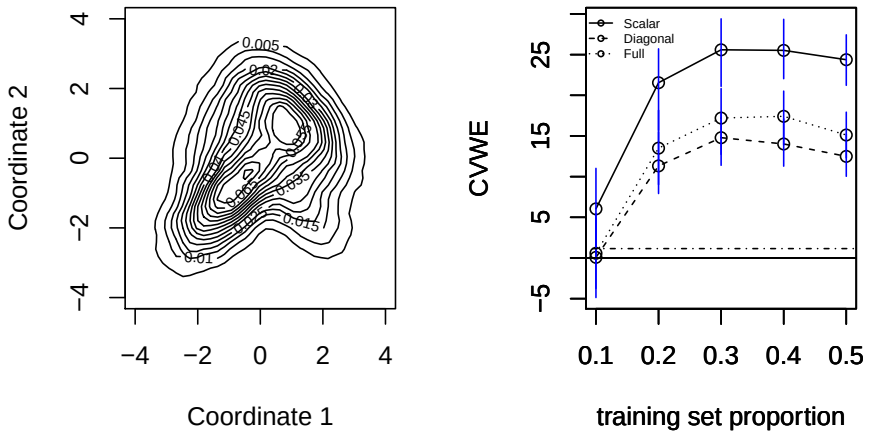

Figure A.12: Testing bivariate normality using $\mathrm{CVWE}_K(\mathcal{S})$, $\mathrm{CVWE}_K(\mathcal{D})$, and $\mathrm{CVWE}_K(\mathcal{F})$ for quadrimodal data.